A map of human cancer signaling

Cui, Qinghua; Ma, Yun; Jaramillo, Maria; Bari, Hamza; Awan, Arif; Yang, Song; Zhang, Simo; Liu, Lixue; Lu, Meng; O'Connor-McCourt, Maureen; Purisima, Enrico O.; Wang, Edwin

Publisher's version / Version de l'éditeur:
https://doi.org/10.1038/msb4100200
Molecular Systems Biology, 3, pp. 152-164, 2007

Molecular Systems Biology 3; Article number 152; doi:10.1038/msb4100200
Citation: ***Molecular Systems Biology* 3:152**
© 2007 EMBO and Nature Publishing Group All rights reserved 1744-4292/07
www.molecularsystemsbiology.com

A map of human cancer signaling

Qinghua Cui[1], Yun Ma[2], Maria Jaramillo[3], Hamza Bari[1], Arif Awan[1], Song Yang[4], Simo Zhang[2], Lixue Liu[2], Meng Lu[2], Maureen O'Connor-McCourt[3], Enrico O Purisima[1,5] and Edwin Wang[1,5,]*

[1] Computational Chemistry and Biology Group, Biotechnology Research Institute, National Research Council Canada, Montreal, QC, Canada, [2] Department of Biology, Tianjin Normal University, Tianjin, China, [3] Receptor, Signaling and Proteomics Group, Biotechnology Research Institute, National Research Council Canada, Montreal, QC, Canada, [4] School of Chemical Engineering, Tianjin University, Tianjin, China and [5] Center for Bioinformatics, McGill University, Montreal, QC, Canada
* Corresponding author. Computational Chemistry and Biology Group, Biotechnology Research Institute, National Research Council Canada, 6100 Royalmount, Montreal, QC, Canada H4P 2R2. Tel.: +1 514 496 0914; Fax: +1 514 496 0943; E-mail: edwin.wang@cnrc-nrc.gc.ca

Received 8.6.07; accepted 6.11.07

We conducted a comprehensive analysis of a manually curated human signaling network containing 1634 nodes and 5089 signaling regulatory relations by integrating cancer-associated genetically and epigenetically altered genes. We find that cancer mutated genes are enriched in positive signaling regulatory loops, whereas the cancer-associated methylated genes are enriched in negative signaling regulatory loops. We further characterized an overall picture of the cancer-signaling architectural and functional organization. From the network, we extracted an oncogene-signaling map, which contains 326 nodes, 892 links and the interconnections of mutated and methylated genes. The map can be decomposed into 12 topological regions or oncogene-signaling blocks, including a few 'oncogene-signaling-dependent blocks' in which frequently used oncogene-signaling events are enriched. One such block, in which the genes are highly mutated and methylated, appears in most tumors and thus plays a central role in cancer signaling. Functional collaborations between two oncogene-signaling-dependent blocks occur in most tumors, although breast and lung tumors exhibit more complex collaborative patterns between multiple blocks than other cancer types. Benchmarking two data sets derived from systematic screening of mutations in tumors further reinforced our findings that, although the mutations are tremendously diverse and complex at the gene level, clear patterns of oncogene-signaling collaborations emerge recurrently at the network level. Finally, the mutated genes in the network could be used to discover novel cancer-associated genes and biomarkers.

Molecular Systems Biology 18 December 2007; doi:10.1038/msb4100200
Subject Categories: metabolic and regulatory networks; molecular biology of disease
Keywords: cancer signaling; cancer-signaling map; DNA methylation; gene mutation; oncogene-signaling dependence; signaling network; tumorigenesis

Introduction

Cells use sophisticated communication between proteins in order to initiate and maintain basic cellular functions such as growth, survival, proliferation and development. Traditionally, cell signaling is described via linear diagrams and signaling pathways. As many more 'cross-talks' between signaling pathways have been identified (Natarajan *et al*, 2006), a network view of cell signaling emerged: the signaling proteins rarely operate in isolation through linear pathways, but rather through a large and complex network. As cell signaling is crucial to affect cell responses such as growth and survival, alterations of cellular signaling events, such as those that arise by mutations, can result in tumor development. Indeed, cancer is largely a genetic disease that is caused by acquiring genomic alterations in somatic cells. Alterations to the genes that encode key signaling proteins, such as RAS and PI3K, are commonly observed in many types of cancers. During tumor progression, it is proposed that a malignant tumor arises from a single cell, which undergoes a series of evolutionary processes of genetic or epigenetic changes and selections so that a cell within the population can acquire additional selective advantages for cellular growth or survival, resulting in progressive clonal expansion (Nowell, 1976).

Genetic mutations of the signaling proteins might over-activate key cell-signaling properties such as cell proliferation or survival and then give rise to the cell with selective advantages for uncontrolled cellular growth and promoting tumor progression. In addition, mutations may also inhibit the function of tumor-suppressor proteins, resulting in a relief from normal constraints on growth. Furthermore, epigenetic alterations by promoter methylation, resulting in transcriptional repression of genes controlling tumor malignancy, is another important mechanism for the loss of gene function that can provide a selective advantage to tumor cells.

Enormous efforts have been made over the past few decades to identify mutated genes that are causally implicated in human cancer. Furthermore, a genome-wide or large-scale sequencing of tumor samples across many kinds of cancers represents a largely unbiased overview of the spectrum of mutations in human cancers (Stephens *et al*, 2005; Sjoblom *et al*, 2006; Greenman *et al*, 2007; Thomas *et al*, 2007). Most of these efforts have been made by the Cancer Genome Project (CGP, http://www.sanger.ac.uk/genetics/CGP/), which aims to identify cancer-mutated genes using genome-wide mutation-detection approaches. Similarly, genome-wide identification of epigenetic changes in cancer cells has been conducted recently (Ohm *et al*, 2007; Schlesinger *et al*, 2007; Widschwendter *et al*, 2007). These studies showed that a substantial fraction of the cancer-associated mutated and methylated genes is involved in cell signaling, which is in agreement with the previous finding that the most common domain encoded by cancer genes is the protein kinase domain (Futreal *et al*, 2004). Although there is a wealth of knowledge regarding molecular signaling in cancer, the complexity of human cancer prevents us from gaining an overall picture of the mechanisms by which these genetic and epigenetic events affect cancer cell signaling and tumor progression. Where are the oncogenic stimuli embedded in the network architecture? What are the principles by which genetic and epigenetic alterations trigger oncogene-signaling events? Given that so many genes have genetic and epigenetic aberrations in cancer signaling, what is the architecture of cancer signaling? Do any tumor-driven signaling events represent ‘oncogenic dependence’ (the phenomenon by which certain cancer cells become dependent on certain signaling cascades for growth or survival)? Who are the central players in oncogene signaling? Are there any signaling partnerships generally used to generate tumor phenotypes? To answer these questions, we conducted a comprehensive analysis of cancer mutated and methylated genes on a human signaling network, focusing on network structural aspects and quantitative analysis of gene mutations on the network.

Results and discussion

The architecture and the relationships among the proteins of a signaling network are important for determining the sites at which oncogenic stimuli occur and through which oncogenic stimuli are transduced. Extensive signaling studies during the past decades have yielded an enormous amount of information regarding regulation of signaling proteins for more than 200 signaling pathways, most of which have been assembled and collected in public databases in diagrams. We manually curated the data on signaling proteins and their relations (activation and inhibitory and physical interactions) from the BioCarta database and the Cancer Cell Map database (see Materials and methods). We merged the curated data with another literature-mined signaling network that contains $\sim$500 proteins (Ma’ayan *et al*, 2005). As a result, we have built a human signaling network containing 1634 nodes and 5089 links. Integrative network analyses have provided numerous biological insights (Wuchty *et al*, 2003; Han *et al*, 2004; Ihmels *et al*, 2004; Luscombe *et al*, 2004; Kharchenko *et al*, 2005; Wang and Purisima, 2005; Cui *et al*, 2006). Thus, the integration of the data on mutated and methylated cancer-associated genes onto the network will help us to identify critical sites involved in tumorigenesis and increase our understanding of the underlying mechanisms in cancer signaling.

To integrate mutated and methylated genes onto the network, we first collected the cancer mutated genes from the Catalogue Of Somatic Mutations In Cancer (COSMIC) database, which collects the cancer mutated genes through literature curation and large-scale sequencing of tumor samples in the CGP. We then combined these data with the cancer mutated genes derived from other genome-wide and high-throughput sequencing of tumor samples (Stephens *et al*, 2005; Sjoblom *et al*, 2006; Greenman *et al*, 2007; Thomas *et al*, 2007). The merged gene set represents a mixture of the past directed approach and current systematic screening of cancer mutations. The cancer-associated methylated genes were taken from the genome-wide identification of the DNA methylated genes in cancer stem cells (Ohm *et al*, 2007; Schlesinger *et al*, 2007; Widschwendter *et al*, 2007). Finally, 227 cancer mutated genes and 93 DNA methylated genes were mapped onto the network. Among the 227 cancer mutated genes, 218 (96%) and 55 (24%) genes were derived from large-scale gene sequencing of tumors and literature curation, respectively (see Materials and methods, Figure 1A). In general, cancer genes can be divided into two groups: positive regulators (oncogenes) that promote cancer cell proliferation and the negative regulators (tumor suppressors) that restrain it. By comparing the mutated genes with the known tumor suppressors, we found that only 6.6% (15 genes) of the mutated genes are known tumor suppressors and that the majority of the mutated genes are oncogenes (Supplementary Figure 1). On the other hand, methylated genes are mainly found to encode tumor suppressors in cancer cells (Supplementary Figure 1) (Ohm *et al*, 2007; Widschwendter *et al*, 2007).

Cancer mutated genes are enriched in signaling hubs but not in neutral hubs

Genes that, when mutated or silenced, result in tumorigenesis often lead to the aberrant activation of certain downstream signaling nodes resulting in dysregulated growth, survival and/or differentiation. The architecture of a signaling network is important for determining the site at which a genetic defect is involved in cancer. To discover where the critical tumor signaling stimuli occur on the network, we explored the network characteristics of the mutated and methylated genes. The signaling network is presented as a graph, in which nodes represent proteins. Directed links are operationally defined to represent effector actions such as activation or inhibition, whereas undirected links represent protein physical interactions that are not characterized as either activating or inhibitory. For example, scaffold proteins do not directly activate or inhibit other proteins but provide regional

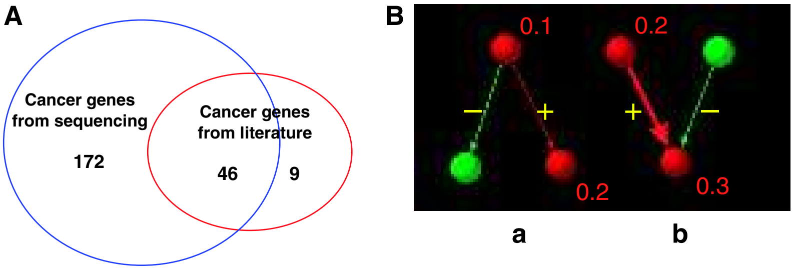

Figure 1 Illustration of the sources of cancer mutated network genes and oncogenic signal transduction events. (**A**) Most of the cancer mutated network genes were discovered by large-scale sequencing of tumor samples, whereas a small fraction of them was found in literature. (**B**) Oncogenic signal transduction events and oncogene-signaling-dependent events. (a) Signaling divergent unit. The line in red represents an oncogenic signal transduction event. (b) Signaling convergent unit. The line in red represents an oncogene-signaling-dependent event. In this case, both genes have high mutation frequency ($\geqslant$0.02), suggesting that the signaling event between the two genes is frequently used in tumorigenesis. Nodes in red represent mutated genes, whereas numbers represent mutation frequencies. Signs $+$ and $-$ represent activating and inhibitory links, respectively.

organization for activation or inhibition between other proteins through protein interactions. In this case, undirected links are used to represent the interactions between scaffold proteins and others. On the other hand, adaptor proteins are able to activate or inhibit other proteins through direct protein interactions. In this situation, directed links are used to represent these relations. There are two kinds of directed links. An incoming link represents a signal from another node. The sum of the number of incoming links of a node is called the indegree of that node. An outgoing link represents a signal to another node. The sum of the number of outgoing links of a node is called the outdegree of that node. We call incoming and outgoing links as signal links, whereas the physical links are neutral links. We first examined the characteristics of the nodes that represent mutated genes on the network. We compared the average indegree of the mutated genes with that of the nodes in the whole network. We found that the average indegree of the mutated nodes is significantly higher than that of the network nodes ($P<1.1\times10^{-6}$, Wilcoxon test, Supplementary Figure 2). A similar result was obtained for the average outdegree of the mutated nodes ($P<6.0\times10^{-14}$, Wilcoxon test, Supplementary Figure 2). In contrast, there is no difference of the average neutral degrees between the mutated nodes and other nodes in the network. To refine these results further, we calculated the correlations between the indegree, outdegree and neutral degree of the network nodes. We found a significant correlation between the indegree and the outdegree of the network nodes (R=0.41, $P<2.2\times10^{-16}$, Spearman's correlation), but no correlation between the indegree and neutral degree of the nodes (R=−0.02, P=0.54, Spearman's correlation). Taken together, these results suggest that cancer mutations most likely occur in signaling proteins that are acting as signaling hubs (i.e., RAS) actively sending or receiving signals rather than in nodes that are simply involved in passive physical interactions with other proteins. As these hubs are focal nodes that are shared by, and/or are central in, many signaling pathways, alterations of these nodes, or signaling hubs, are predicted to affect more signaling events, resulting in cancer or other diseases. In previous studies, we found that cancer-associated genes are enriched in hubs (Awan *et al*, 2007). However, these results indicate that cancer-associated genes are enriched in signaling hubs but not neutral hubs.

We also investigated the relations between the node degree and the methylated genes in the network. Methylated gene nodes do not appear to differ significantly from the network nodes with regard to their indegree, outdegree and neutral degree, respectively (P=0.32, P=0.16, P=0.09, Supplementary Figure 2). These results suggest that cancer mutated genes and methylation-silenced genes have different regulatory mechanisms in oncogene signaling.

Activating and inhibitory signals enhance and alleviate oncogene-signaling flows, respectively

Signaling flow branching represents the splitting of one signal at a source node (Figure 1B), whereas signaling flow convergence represents the consolidation of the signals at a target node from two source nodes (Figure 1B). Both types of the signaling flows are the basic elements of the network architectural organization. In the network, when the upstream and downstream nodes of a particular signal transduction event get altered either genetically or epigenetically, we considered the transduction event (link) to be most likely selected and used in cancer signaling and defined it as an oncogenic signal transduction event (Figure 1B). If a particular oncogenic signal transduction event is frequently found in many tumor samples, we infer that the tumor cells are 'dependent' on this highly used signaling event and call it 'oncogene-signaling-dependent event' (Figure 1B). To investigate how cancer signaling is distributed on these signal transduction routes, we extracted all the branching and convergent signaling flow units that contain at least one oncogenic signal transduction event and conducted a quantitative analysis by overlaying the gene mutation frequency onto these units. The mutation frequency of a gene was defined as the number of tumor samples that contain that mutated gene divided by the total number of the tumor samples that are used to screen the mutations for that gene. The mutation frequency of each mutated gene was obtained by using the COSMIC database, which contains the data on more than 200 000 tumor samples screened for cancer gene mutations. For the signaling branching units, we divided the signaling flows into two groups: activating and inhibitory group (Figure 1B) and compared the gene mutation frequencies of the upstream

Table I Effects of the positive and negative signals on the oncogene-signaling flows

	Signaling branching type		Signaling convergence type	
	Increasing	Decreasing	Increasing	Decreasing
Activating group	676	551	1032	418
Inhibitory group	46	140	93	96
Odds ratio	3.7		2.5	
P-value	3.7×10^{-15}		2.5×10^{-9}	

For each signaling flow type, we classified the units into two groups: activating group and inhibitory group. For each group, we compared the mutation frequencies of an upstream node with that of the downstream node. 'Increasing' represents that the mutation frequency of a downstream node is higher than that of an upstream node, whereas 'decreasing' represents that the mutation frequency of a downstream node is lower than that of an upstream node. Odds ratio was calculated by (increasing, activating group) × (decreasing, inhibitory group)/(increasing, inhibitory group)/(decreasing, activating group). *P*-value was calculated by Fisher's exact test.

nodes with those of the downstream nodes in each group. Interestingly, in the activating group, the upstream nodes often have lower mutation frequencies than those of the downstream nodes. In contrast, in the inhibitory group, the upstream nodes often have higher mutation frequencies than those of the downstream nodes (Table I). Statistical tests confirmed that these observations are statistically significant (Table I). Similar results were obtained for the signaling convergent units as well (Figure 1B, Table I). These results suggest that the oncogene-signaling event triggered by mutations is preferentially associated with activating downstream signaling paths or conduits. Conversely, oncogene-signaling event triggered by mutations are less likely to be associated with downstream inhibitory signaling paths.

In general, there are far more activating signaling flows than inhibitory ones in the network. Thus, we hypothesized that the downstream genes of the network, especially the genes of the output layer of the network, would have a higher mutation frequency. To test this possibility, we compared the average gene mutation frequency of the nuclear proteins, which represent the output layer members of the network, with that of the other network genes. Indeed, the nuclear genes have higher mutation frequency than others (P=0.01, Wilcoxon test), which complements with our previous finding that cancer-associated genes are enriched in the nuclear proteins (Awan *et al*, 2007). In contrast, the distributions of the methylated genes have no such preference, suggesting that DNA methylated genes do not tend to directly affect the output layer of the network. These results strongly suggest that the genes in the output layer of the network, which play direct and important roles in determining phenotypic outputs, are frequent targets for activating mutations. The importance of this output layer is reinforced by our previous observations that the expression of the output layer genes of the signaling network is heavily regulated by microRNAs (Cui *et al*, 2006).

Mutated and methylated genes are enriched in positive and negative regulatory loops, respectively

The complex architecture of signaling networks can be regarded as consisting of interacting network motifs, which are statistically overrepresented subgraphs that appear recurrently in networks. A signaling network motif, also known as regulatory loops in biology, is a group of interacting proteins capable of signal processing. They bear specific regulatory properties and mechanisms (Babu *et al*, 2004; Wang and Purisima, 2005). The structure and the intrinsic properties of the frequently occurring network regulatory motifs give us a functional view of the organization of signaling networks. Thus, the study of the distributions of the mutated and methylated genes in the network motifs will provide insights into cancer-signaling regulatory mechanisms. We first examined the mutated genes on the feed-forward loops, in which the first protein regulates the second protein, and both proteins regulate the third protein. We classified the feed-forward loops into four subgroups (labeled 0–3) based on the number of nodes that are mutated genes. We calculated the ratio (Ra) of positive (activating) links to the total directed (positive and negative) links in each subgroup and compared it with the average Ra in all the feed-forward loops, which is shown as a horizontal line in Supplementary Figure 3. The Ra ($\sim$0.7) in subgroup 0 is less than the average Ra ($\sim$0.74) of all the feed-forward loops ($P < 1.9 \times 10^{-9}$, Fisher's test). However, as the number of mutated nodes rises, the Ra for the corresponding group increases to a maximum of $\sim$0.93 (Supplementary Figure 3, Supplementary Table 1). We obtained similar results, when we extended the same analysis to all the 3-node- and 4-node-size network motifs (Figure 2, Supplementary Table 1). These motifs show a clear positive correlation between positive link ratio and the number of mutated genes in the motifs. These results suggest that cancer gene mutations occur preferentially in positive regulatory motifs. In contrast, all the 3-node and 4-node size motifs show an obviously negative correlation between positive link ratio and the number of methylated genes in the motifs (Figure 2, Supplementary Table 2). These results suggest that cancer gene methylation preferentially occurs in negative regulatory motifs. A similar trend was found for the 15 known tumor suppressors (Supplementary Figure 4a–d), which is in agreement with the notion that cancer-associated methylated genes play roles as tumor suppressors. Collectively, these facts suggest that mutated and methylated genes have different regulatory mechanisms in cancer signaling and support the notion that gene mutations and methylations are strongly selected in tumor samples.

Signaling information propagates through a series of built-in regulatory motifs to contribute to cellular phenotypic functions (Ma'ayan *et al*, 2005). The transition from a normal cellular state into a long-term deregulated state such as cancer is often driven by prolonged activation of downstream proteins, which are regulated by upstream proteins or regulatory motifs or circuits. Positive regulatory loops (Ferrell, 2002) could amplify signals, promote the persistence of signals, serve as information storage and evoke biological responses to generate phenotypes such as cancer. In cancer

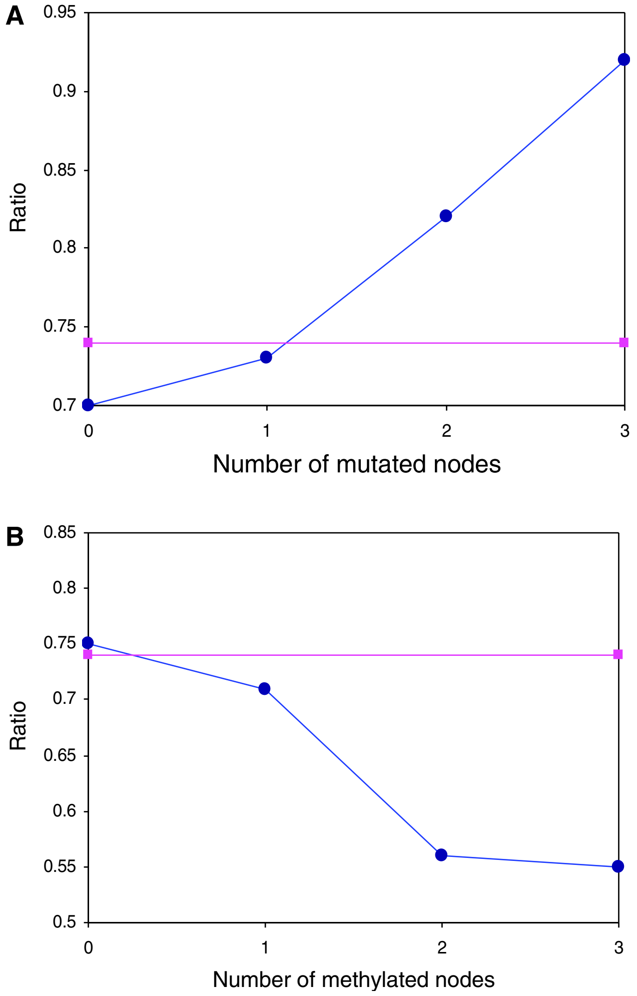

Figure 2 Enrichment of mutated and methylated genes in network motifs. (**A**) Relations between the fractions of positive links in all 3-node-size network motifs and the fractions of mutated genes in these motifs. (**B**) Relations between the fractions of positive links in all 3-node-size network motifs and the fractions of methylated genes in these motifs. All network motifs were classified into subgroups based on the number of nodes that are either mutated genes or methylated genes, respectively. The ratio of positive links to total positive and negative links in each subgroup was plotted. The horizontal lines indicate the ratio of positive links to the total positive and negative links in all network motifs.

cells, constitutive activation of the oncogene signaling is necessary. Neutral mutations do not affect protein function, whereas missense mutations may have positive or negative effects on protein activity. The enrichment of gene mutations in positive regulatory loops suggests that the mutants in the motifs must have gain of function or increase their biochemical activities compared to the wild-type genes in order to constitutively activate downstream proteins. Indeed, a recent survey showed that 14 out of the 15 PI3K mutants in tumors have gain of function (Gymnopoulos *et al*, 2007). Gain-of-function mutants in a positive regulatory loop afford the amplification of weak input stimuli and serve as information storage to extend the duration of activation of the affected downstream proteins. This might allow the downstream signaling cascades to persistently hold and transfer information leading to tumor phenotypes.

Promoter gene methylation is a known mechanism of inducing loss of function by inhibiting the expression of genes (Ohm *et al*, 2007; Widschwendter *et al*, 2007). Negative regulatory loops controlled by tumor-suppressor proteins repress positive signals and play an important role in maintaining cellular homeostasis and restraining the cellular state transitions (Ma'ayan *et al*, 2005). A loss of function of gene methylation in a negative regulatory loop could break the negative feedbacks, thereby releasing the restrained activation signals and promoting oncogenic state transitions. Homeostasis relies on the balance between positive and negative signals in crucial components of the network. Both the gain-of-function mutated genes in positive regulatory loops and the loss-of-function methylated genes in negative regulatory loops could break this delicate balance, thus promoting state transitions and generating tumor phenotypes. Therefore, both mutated and methylated genes and their regulatory loops (oncogenic regulatory loops) are critical components of the network where the oncogenic stimuli occur.

An oncogene-signaling map emerges from the network

In the language of networks, genes whose mutations or epigenetic silencing are crucial to trigger oncogene signaling might link together as components in the network. Identification of such components will help us to discover the relationships and structural organizations of the oncogenic proteins. To uncover the architecture of cancer signaling and to gain insights into the higher-order regulatory relationships among signaling proteins that govern oncogenic signal stimuli, we mapped all genetic mutations and epigenetically silenced genes onto the network. We found that most of these genes (67%) are connected together to form a giant, linked network component. Randomization tests confirmed that such a component is unlikely to be formed by chance ($P<2\times10^{-4}$). To build an oncogenic map, we included other mutated and methylated genes that are not present in the composition of the component into the giant network component based on node connectivity (see Materials and methods). The resulting oncogene-signaling map consists of approximately 20% of the signaling network nodes (326 nodes, 892 links) and includes almost 90% of the mutated and methylated genes (Figure 3). The map showed different network topological characteristics from the signaling network. For example, the average length of the map is less than that of the signaling network (5 versus 6, $P<2\times10^{-16}$, Wilcoxon test). On the other hand, the average clustering coefficient of the map is greater than that of the signaling network (0.08 versus 0.04, P=0.06, Wilcoxon test). These results suggest that oncogenic proteins tend to have more interactions and signaling regulatory relationships. The emerging oncogene-signaling map represents a 'hot area' where extensive oncogene-signaling events might occur. As a proof of concept, we found that the MAPK kinase and TGFβ pathways, which are well-known cancer-signaling pathways, are embedded in the map. For example, 50 out of 87 proteins in the MAPK kinase pathway (Supplementary Table 3) and 22 out of 52 proteins in the TGFβ pathway (Supplementary Table 4), respectively, are

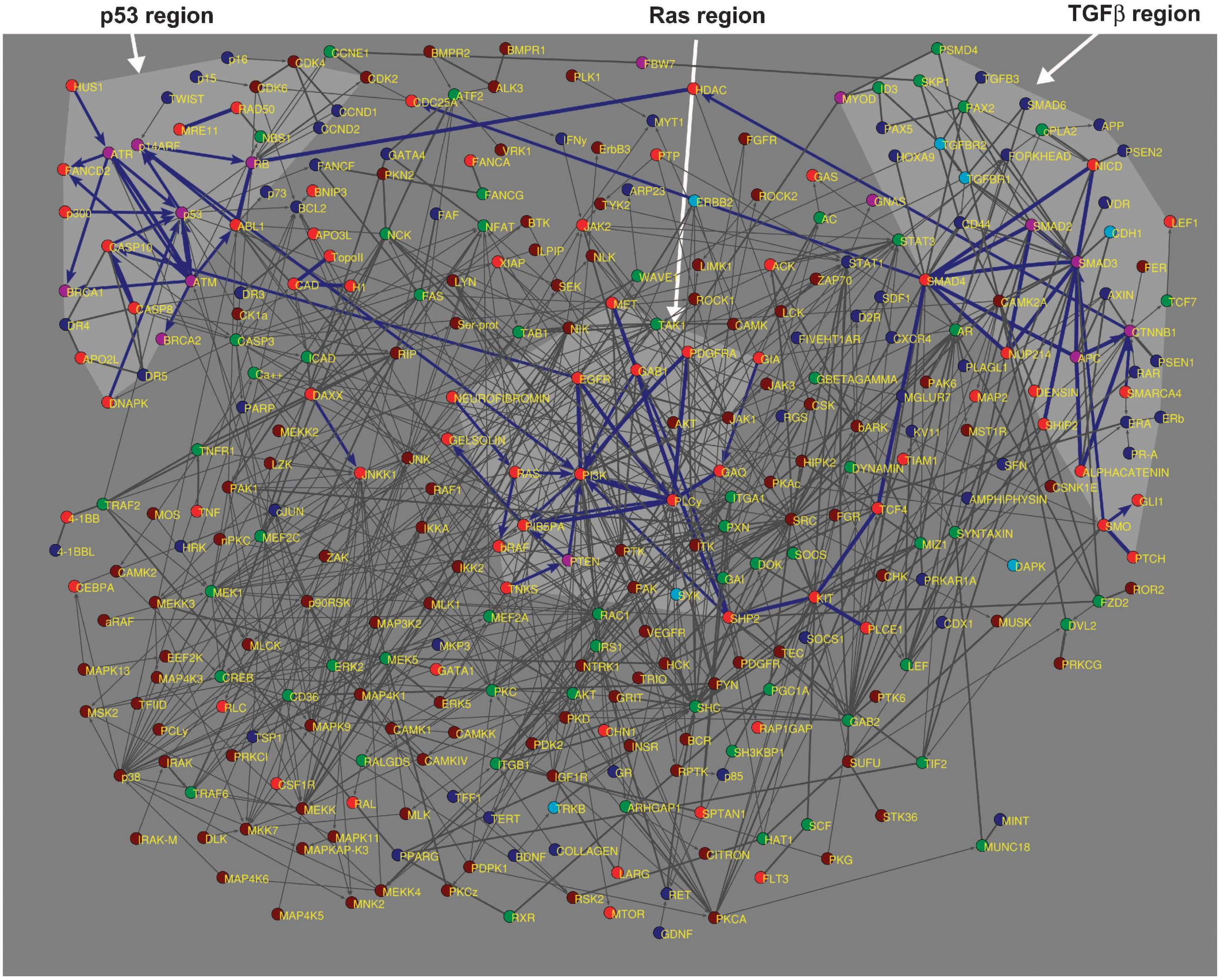

Figure 3 Human oncogene-signaling map. The human cancer-signaling map was extracted from the human signaling network, which was mapped with cancer mutated and methylated genes. The map shows three 'oncogenic-dependent regions' (background in light gray), in which genes of the two regions are also heavily methylated. Nodes represent genes, whereas the links with and without arrows represent signal and physical relations, respectively. Nodes in red, purple, brown, cyan, blue and green represent the genes that are highly mutated but not methylated, both highly mutated and methylated, poorly mutated but not methylated, both poorly mutated and methylated, methylated but not mutated, and neither mutated nor methylated, respectively.

included in the map. More importantly, in addition to known oncogenic pathways, there are many other novel candidate cancer-signaling cascades present in the map. For a given gene mutation in a tumor, one could use this map to generate testable hypotheses to discover the underlying oncogene-signaling cascades in that tumor.

As mentioned above, oncogene-signaling-dependent events, which we define as the interactions between the cancer mutated or methylated genes, are frequently found in tumor samples and represent various oncogene-driving events that could play more critical roles in generating tumor phenotypes. To systematically identify such events and discover how they are organized in the map, we charted the gene mutation frequency onto the map and highlighted the signaling links between any two genes that have high mutation frequencies. Most genes have mutation frequencies lower than 2%, whereas a handful of genes have very high mutation frequencies, such as p53 (41%), PI3K (10%) and RAS (15%) (see Materials and methods). Therefore, a gene mutation frequency equal to or greater than 2% was considered as high. Interestingly, nearly 10% of the links in the map are oncogene-signaling-dependent events. Certain signaling events such as Pten-PI3K and RAS-PI3K in the map are well-known oncogene-signaling-dependent events/cascades that are frequently used in various cancers.

As shown in Figure 3, most oncogene-signaling-dependent events are connected, and three major regions that contain densely connected oncogene-signaling-dependent events emerge in the map: the first region (p53 region) contains mainly tumor suppressors such as p53, Rb, BRCA1, BRCA2 and p14 (CDKN2A) etc.; the second region (RAS region) contains mainly well-known oncogenes such as RAS, EGFR and PI3K etc.; and the third region (TGFβ region) contains SMAD3, SMAD4 and a few other TGFβ-signaling proteins. Interestingly, genes in the p53 and TGFβ regions are also heavily methylated in cancer stem cells, suggesting that these regions are involved in the early stage of oncogenesis. Other methylated genes are intertwined with the mutated genes in the map, suggesting that they share some oncogene-signaling cascades and might be regulated to cooperate in cancer signaling via gene mutation and/or methylation. Notably, it seems that, in cancer stem cells, TGFβ-signaling pathway is shut down, supporting its known role as a tumor suppressor in the early stages of tumorigenesis (Hanahan and Weinberg, 2000; Siegel and Massague, 2003). These results suggest that the crucial players of oncogene signaling tend to be closely clustered and regionalized. This map uncovers the architectural structure of the basic oncogene signaling and highlights the signaling events that are highly conserved in generating tumor phenotypes.

Functional collaboration of genes between oncogene-signaling blocks

The oncogene-signaling map can be decomposed into several network communities or network themes (Zhang *et al*, 2005), in which each network community contains a set of more closely linked nodes and ties to particular biological functions. To discover such network communities, we implemented and applied an algorithm that detects network communities to the map. As a result, 12 network communities, ranging in size from 11–65 nodes (Supplementary Table 5), called 'oncogene-signaling blocks', were found in the map. Structurally, the nodes within each block have more links and signaling regulatory relations among themselves than others. The genes in each block share similar biological functions such as cell proliferation, development and apoptosis (Supplementary Table 5). We further performed Gene Ontology (GO) enrichment analysis for each oncogene-signaling block using DAVID Tools (http://david.abcc.ncifcrf.gov/home.jsp). Most of the oncogene-signaling blocks are enriched with protein serine/threonine kinase activity (Supplementary Table 6), which is well known to take part in tumorigenesis. Notably, Block 1 is enriched with cell surface receptor-linked signaling, whereas Block 10 is enriched with intracellular signaling cascades. Block 11 is enriched with tumor suppressors and biological processes such as apoptosis and cell cycle. These results suggest that certain blocks are taking part in different parts/kinds of signaling, that is, cell surface receptor-related signaling, intracellular signaling, cascade signaling and apoptotic signaling. However, three oncogene-signaling blocks have no GO enrichment detected. One of the reasons is that a fraction of the genes in these blocks is not well annotated yet. For example, about one-third of genes in Block 6 have no GO term associated.

We asked if the genes in each block could operate in a compensatory or concerted manner to govern a set of similar functions. Toward this end, we surveyed the gene mutations in tumor samples where at least two genes are screened for mutations. As a result, the co-occurrence in tumor samples of 25 mutated gene pairs is found to be statistically significant (Supplementary Table 7). Significantly, only three collaborative gene pairs came from the same block, whereas other collaborative gene pairs came from two different blocks, with predominantly one of them arising from Block 11 (defined as p53 block), which contains p53, Rb, p14, BRCA1, BRCA2 and several other genes involved in control of DNA damage repair and cell division. Collectively, these results suggest that the signaling genes from different blocks most likely work together in a complementary way to generate tumor phenotypes.

We further asked which oncogene-signaling blocks work together to produce a tumor phenotype. To address this question, we surveyed the gene mutations in the tumor samples where at least two gene mutations are found. In total, 592 tumor samples fit this criterion. We used the 592 samples to build a matrix (M) where samples are rows and the signaling blocks are columns. If a gene of a particular signaling block (b) gets mutated in a tumor sample(s), we set $M_{\mathrm{s,b}}$ to 1, otherwise we set $M_{\mathrm{s,b}}$ to 0. A heatmap was generated using the matrix (Figure 4A). If a sample contains statistically significant co-occurring mutated gene pairs (see Supplementary Table 7), these pairs were highlighted in the heatmap. Samples were organized based on the cancer types they belong to. Several cancer types such as breast, central nervous system, blood, lung, pancreas and skin tumors that have relatively more samples were also highlighted in the heatmap. As shown in Figure 4A, two signaling blocks have statistically significant enrichment of gene mutations ($P<2\times10^{-4}$, randomization tests), suggesting that genes in these two signaling blocks are predominantly used to generate tumor phenotypes. One oncogene-signaling block (Block 1, defined as RAS block) contains genes like RAS, EGFR and PI3K etc., which share similar biological functions such as cell proliferation, cell survival and cell growth, whereas the other is the p53 block, which share similar biological functions such as cell cycle checkpoint control, apoptosis and affecting genomic instability (Supplementary Table 5). These two blocks also represent the two oncogene-signaling-dependent regions (p53 and RAS regions) in Figure 3, respectively. When a tumor sample has a mutation in a gene from the RAS-signaling block, it is also most likely to contain a mutation in a gene from the p53 block ($P<2\times10^{-4}$). To check if this phenomenon is primarily due to a particular pair of genes, we calculated the likelihood of co-occurrence for each pair of the genes, of which one gene is mutated in one block and the other gene is mutated in the other block. We found that the *P*-values for gene pairs are always significantly greater than that for the pair of Blocks RAS and p53. For example, the *P*-value of co-occurrence of RAS (in Block RAS) and p53 (in Block p53) mutations is 0.01, which is greater than that of the two blocks ($P<2\times10^{-4}$). This indicates that these two oncogene-signaling blocks collaborate to generate tumor phenotypes for most tumors. Experimental examples have shown similar gene collaboration in tumorigenesis: activation of RAS (RAS block) and inactivation of p53 (p53 block) induce lung tumors (Meuwissen and Berns, 2005), whereas activation of RAS (RAS block) and inactivation of p16 (p53 block) induce pancreatic tumors (Obata *et al*, 1998). In general, tumor cells exhibit either elevated cell proliferation or reduced differentiation or apoptosis relative to normal cells. The oncogenic blocks we have identified, especially the RAS and p53 blocks, encode functions that are tumor-related, such as cell cycle control, cell proliferation and apoptosis

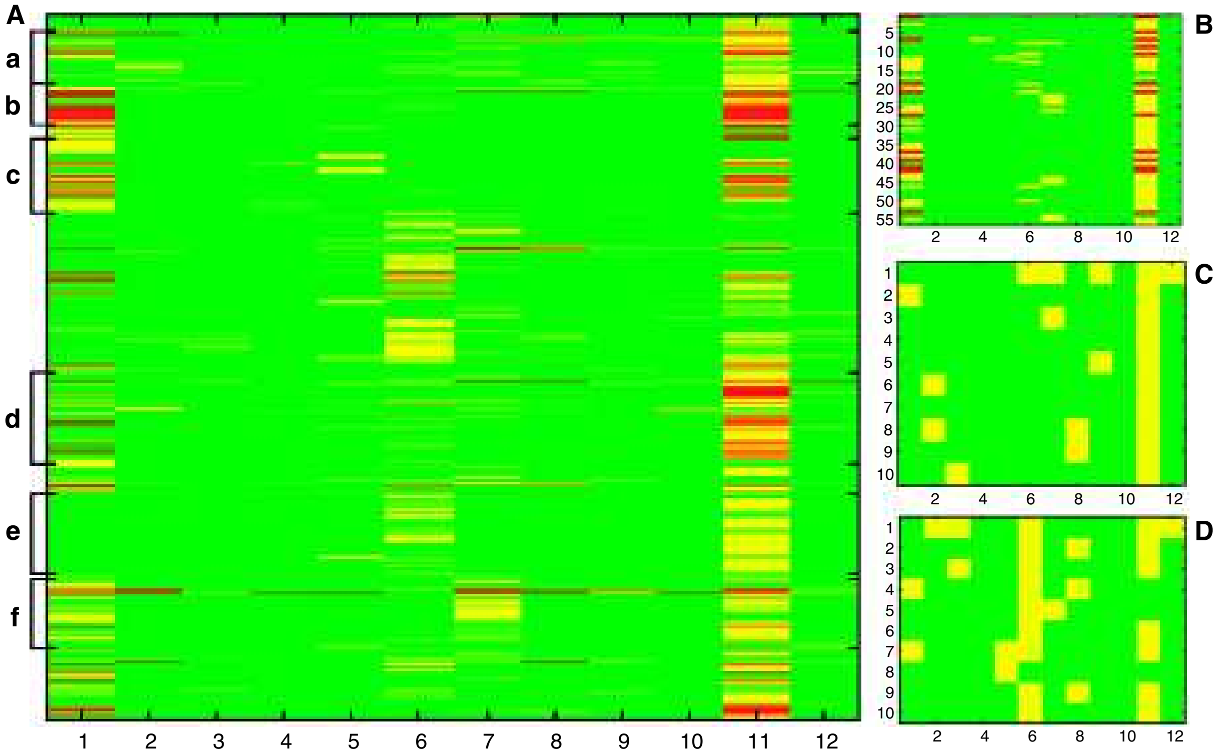

Figure 4 Heatmaps of the gene mutation distributions in oncogene-signaling blocks. Twelve topological regions or oncogene-signaling blocks have been identified based on the gene connectivity of the human oncogene-signaling map. A heatmap was generated from a matrix, which was built by querying the oncogene signaling blocks using tumor samples, in which each sample has at least two mutated genes. If a gene of a particular signaling block (b) gets mutated in a tumor sample (s), we set $M_{s,b}$ to 1, otherwise we set $M_{s,b}$ to 0. (**A**) A heatmap generated using the gene mutation data of the 592 tumor samples. (**B**) A heatmap generated using the gene mutation data of the NCI-60 cancer cell lines. (**C**, **D**) Heatmaps generated using the output from the genome-wide sequencing of breast and colon tumor samples, respectively. Rows represent samples, whereas columns represent oncogene-signaling blocks. Samples were organized according to the cancer types they belong to. Cancer types that have relatively more samples were marked on the heatmap: (a) breast, (b) central nervous system, (c) blood, (d) lung, (e) pancreas and (f) skin tumors. Blocks with gene mutations are marked in yellow; however, when one sample contains statistically significant co-occurring mutated gene pairs (see Supplementary Table 7), the blocks are marked in red.

(Supplementary Figure 5). Activation of genes in the RAS block promotes the cell proliferation, whereas inactivation of genes in the p53 block prevents apoptosis. Thus, a functional collaboration between the genes in these two blocks would promote synergistic cancer signaling and foster tumorigenesis.

Notably, we found that at least one gene mutation in the p53 block had occurred in the tumor samples we examined. In other words, the p53 block is involved in generating tumors for most cancers. This result suggests that the p53 block is a central oncogene-signaling player and essential in tumorigenesis. This finding is further supported by the following observations. (a) To become oncogenic, tumor suppressors require loss-of-function mutations, which occurs more often than gain-of-function mutations (Gymnopoulos *et al*, 2007). Indeed, the average gene mutation frequency in the p53 block is higher than that of other signaling blocks including the RAS block. (b) The methylation of genes in the cancer stem cells resulting in long-term loss of expression represents the early stage of the tumorigenesis. In fact, most of the members of the p53 block are methylated in cancer stem cells. These facts further support that the p53 block might play an important role in the earlier stages of oncogenesis. (c) Gene methylation or inactivating mutations of DNA damage checkpoint genes such as p53 induce genome instability and thus increase the chance of other gene mutations, including the genes of other oncogene-signaling blocks that could functionally collaborate with the p53 block genes to generate tumor phenotypes.

Using the map as a framework, we benchmarked the mutated genes in the NCI-60 cell lines, which represent a panel of well characterized cancer cell lines and various cancer types. A systematic mutation analysis of 24 known cancer genes showed that most NCI-60 cell lines have at least two mutations among the cancer genes examined (Ikediobi *et al*, 2006). We built a matrix and constructed a heatmap using these cell lines and their mutated genes as described above (Figure 4B). Overall, the pattern obtained from the NCI-60 panel resembles that of the 592 tumor panel with both RAS and p53 blocks enriched with gene mutations and exhibiting statistically significant collaborations in these cell lines (Figure 4B, $P<2\times10^{-4}$), which is in agreement with the earlier observations. We also benchmarked the mutated genes derived from a genome-wide sequencing of 22 tumor samples (Sjoblom *et al*, 2006). Among these 22 samples, 10 breast and 10 colon tumor samples have at least two gene mutations in the map. As shown in Figure 4C and D, the p53 block is enriched with gene mutations. For the 10 colon tumor samples, collaboration between Block 6 and Block p53 is established, but for the 10 breast tumors, collaborative patterns between multiple blocks emerged.

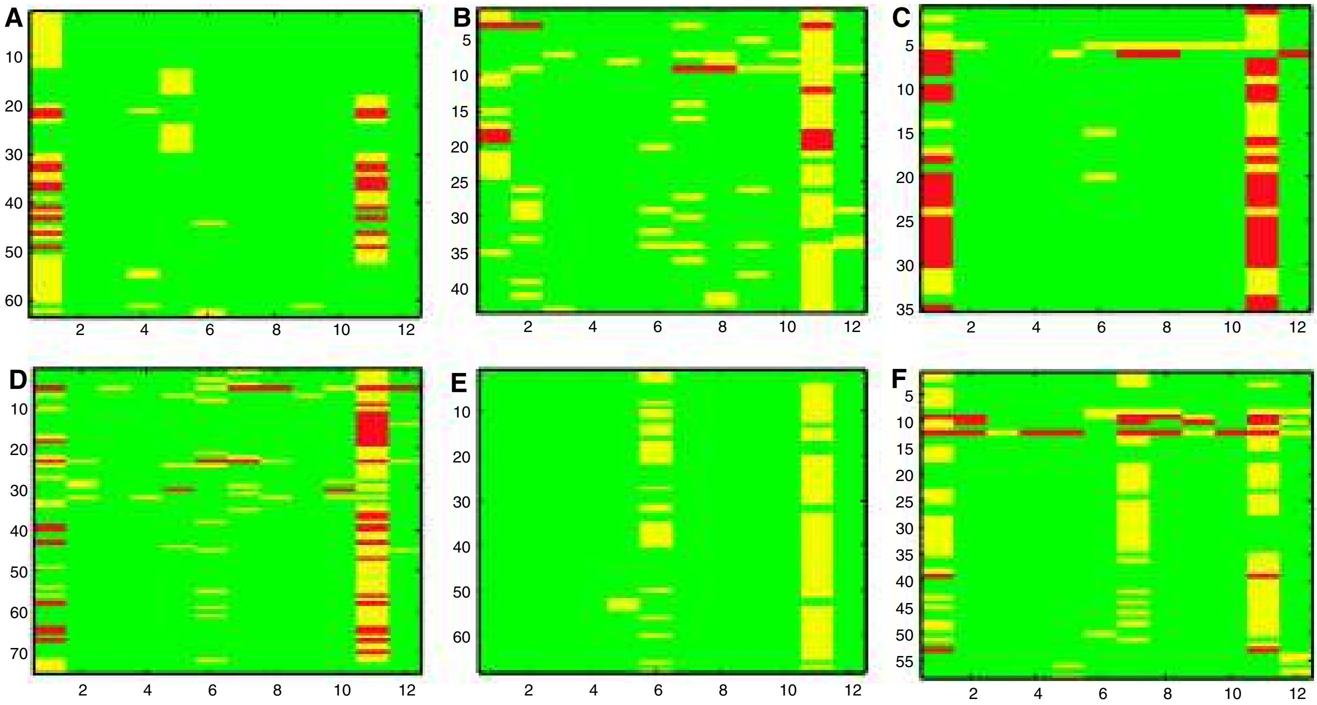

Figure 5 Heatmaps of the gene mutation distributions in oncogene-signaling blocks for six representative cancer types. Heatmaps for (**A**) blood, (**B**) breast, (**C**) central nervous system, (**D**) lung, (**E**) pancreas and (**F**) skin tumors were built using tumor samples of these cancer types, respectively. Rows represent samples, whereas columns represent oncogene-signaling blocks. Blocks with gene mutations are marked in yellow; however, when one sample contains statistically significant co-occurring mutated gene pairs (see Supplementary Table 7), the blocks are marked in red.

To further examine the block collaborative patterns in individual tumor types in higher resolutions, from the heatmap (Figure 4A) we extracted the sub-heatmap for several tumor types that are better represented among the 592 tumor samples, that is, they have relatively more samples within the 592 samples (Figure 5). As shown in Figure 5, signaling block collaborative patterns are tissue dependent and are classified into two groups. One group contains pancreas, skin, central nervous system and blood tumors that have simple block collaborative patterns. In these tumors, signaling collaborations are mainly between Block p53, Block RAS with some minor contributions from Blocks 5, 6 or 7, suggesting that they predominantly use these oncogene-signaling routes to generate tumors resulting in relatively homogenous cancer cell types. The other group contains breast and lung tumors that also contain large proportions of mutations from the p53 block, but also have complex patterns of collaborations between assortments of multiple blocks, suggesting that these tumors may have a larger variety of oncogene-signaling routes, which may explain, in part, the heterogeneous nature of the tumor subtypes in this category. These results might also explain why both lung and breast cancers are the most common types of human tumors.

In this study, the cancer mutated genes were collected from a 'directed approach' (i.e., mutational analysis of specific genes, such as p53) and a 'large-scale approach' (i.e., large-scale sequencing of tumor samples). We tested whether the mutated genes from the directed approach introduce bias to our analysis. Literature-curated cancer mutated genes (directed approach) have been assembled in the Cancer Gene Census (Futreal *et al*, 2004), of which 115 genes were found in the human signaling network. As of November 30th, 2006, among the 115 Cancer Gene Census genes, mutations in 55 of them have been further validated by additional experimental evidence (i.e., other independent experiments confirming the mutation of these specific genes in cancer samples have been documented in the COSMIC database), whereas 60 of them have no such evidence in the COSMIC database (see Materials and methods). In fact, we included only these 55 literature-curated genes in the cancer mutated gene set (227 genes) used in all of our analyses above (see Materials and methods). Of the 55 literature-curated genes, only 9 were not already present in the output of large-scale sequencing of tumor samples (Figure 1A). We removed these 9 genes from the cancer mutated gene set (227 genes), mapped rest of the genes onto the human signaling network and rebuilt an oncogene-signaling map, oncogene-signaling blocks and a heatmap. On the other hand, we added the 60 literature-curated genes, which have no independent supporting evidence in the COSMIC database, to the 227-gene set and obtained 287 genes. Using these 287 genes, we reran the analyses mentioned above. In these two analyses, although the gene members of each oncogene-signaling block have some minor differences with those of the original blocks, the major collaboration patterns of oncogene-signaling blocks remain largely unchanged (Supplementary Figure 6a and b), suggesting that our findings are robust to addition or removal of the cancer mutated genes derived from the directed approach.

The mutated genes in the network provide a predictive power

A substantial number (∼20%) of mutated genes were found in the network. We asked if a gene that has more links to mutated genes in the network is most likely to be cancer associated. To answer this question, we extracted the nonmutated genes that have more than one link to the mutated genes and then grouped them based on their link numbers to the mutated genes. We interrogated a cancer-associated gene set (Supplementary Table 8) compiled from literature mining (see Materials and methods) to find out how many genes in each group are cancer associated. As shown in Figure 6, the more mutated genes a gene links to, the more probably it is cancer associated. When the link number of the network genes is more than six, ∼80% of them are cancer-associated genes. For example, SHC, a gene that has been implicated in cancer metastasis (Jackson *et al*, 2000), has numerous links to the mutated genes in the network. To further investigate the predictive power of the mutated genes in the network, we took the 14 network genes, which not only have at least four links to the mutated genes, but also are not implicated in cancer in the literature, to perform a survival analysis using a microarray data set that contains the gene expression profiles and survival information for 295 breast tumor samples. As a result, the expression profiles of 5 out of the 14 genes (36%) are able to discriminate 'good' and 'bad' tumors (i.e., patients having 'bad' tumors have higher chance of tumor recurrence and short survival time). Therefore, these genes are potentially novel biomarkers. In contrast, less than 10% of the non-mutated network genes have similar discriminatory power. These results suggest that the network genes, which have more links to the mutated genes, have more chance to be perturbed in tumorigenesis and be associated with cancer. Practically, the mutated genes in the network provide a predictive power that can be used to discover novel biomarkers of tumors.

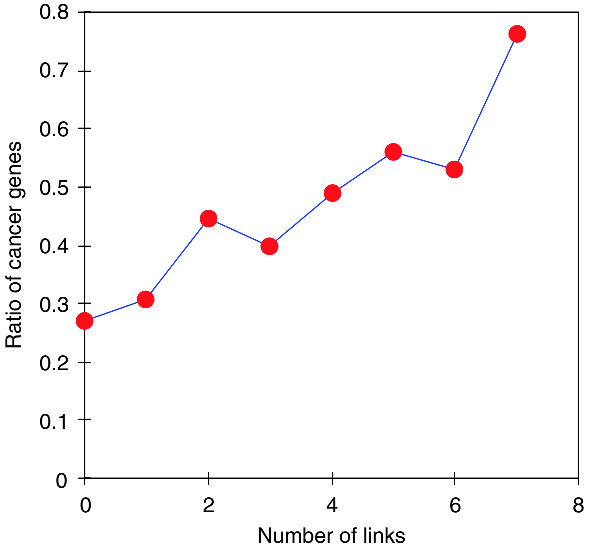

Figure 6 Correlation between the link number of a gene to the mutated genes and cancer-associated genes. We first classified the network genes (without mutations) into groups based on the number of the cancer mutated genes a gene links to. We then calculated the ratio of the cancer-associated genes to total genes for each group. The correlation between the ratio and the groups was plotted.

Concluding remarks

Although a wide variety of genetic and epigenetic events contribute to the signaling of tumorigenesis, it has been challenging to gain a global view of where and how they affect the signaling alterations to generate tumors on the entire signaling network. By integrative analysis of the human signaling network with cancer-associated mutated and methylated genes, we uncovered an overall picture of the network architecture where the oncogenic stimuli occur and the regulatory mechanisms involving mutated and methylated genes. Mutations, the majority of which are activating, preferentially occur in the signaling hub genes (but not neutral hubs) and the genes of the positive regulatory loops, whereas methylated alterations tend to occur in the genes of the negative regulatory loops. Cancer and cell signaling have been well established, and extensive efforts have been made to illustrate cancer signaling during the past few decades. However, it has been a struggle to get clues of how the oncogene signaling is structurally and functionally organized. In this analysis, we extracted an oncogene-signaling map, which provides a blueprint of the oncogene signaling in cancer cells. From the map, we discerned that the oncogene-signaling-dependent events form three highly connected regions that resemble oncogene-signaling superhighways frequently used in tumorigenesis. Topologically, the map has been divided into 12 oncogene-signaling blocks. Functional collaborations between subsets of these blocks are underlying tumorigenesis. In most tumors, genes in both p53 and RAS blocks often get mutated, although the combinations of p53 with other signaling blocks are also found in a small fraction of tumors. Analysis of the NCI-60 cell line panel mutations showed the enrichment of gene mutations in p53 and RAS blocks, which is similar to the patterns found in the 592 samples. Furthermore, we can dissect some of this functional collaboration among different tumor types. These facts indicate that at least two signaling gene mutations, one from the p53 block and the other from another block, are necessary for tumorigenesis. This fact supports the notion that both the prevention of cell death (p53 block) and the promotion of cell proliferation (RAS or other blocks) are necessary to generate most tumors.

At present, a number of researchers doubt or even argue against the value of large-scale human cancer genome sequencing as a meaningful or efficient strategy in cancer research. Their arguments are based on the following observations (Chng, 2007): (a) previous large-scale human cancer genome sequencing revealed that each tumor has a different mutation pattern, and the prevalence and patterns of somatic mutations in human cancers are tremendously diverse and complex (Kaiser, 2006; Sjoblom *et al*, 2006; Greenman *et al*, 2007); (b) the interpretation of such complex somatic alterations is a formidable challenge (Chanock and Thomas, 2007; Thomas *et al*, 2007). We mapped the mutation data from the genome-wide sequencing tumor samples (Sjoblom *et al*, 2006) using the oncogene-signaling map as a framework. Although the number of mutated genes is impressive *in toto*, most signaling gene mutations are limited to 2–3 critical mutations, divided among several signaling blocks, per individual tumor. This result suggests that the mutations in the samples of the same tumor type might share a similar underlying signaling mechanism, because each oncogene-signaling block contains a set of genes linked together through shared regulatory relations and key input and/or output signaling nodes that are involved in tumorigenesis. These findings imply that although the mutations seem tremendously diverse and complex at the gene level, clear patterns emerge recurrently at the network level in most tumors. Therefore, with proper bioinformatics analysis, large-scale cancer genome sequencing efforts would be fruitful in finding appropriate combinations of biological targets for cancer diagnostic and therapeutics.

In summary, this work revealed novel insights into the oncogenic regulatory mechanisms, oncogene-signaling network architecture and oncogene-signaling cooperative relationships that drive cancer development and progression. It also highlights the emergence of the central players in cancer signaling. Cancer studies have integrated microarray, knowledge, pathways and networks (Liu and Lemberger, 2007), but not genetic and epigenetic data yet. However, as the next generation of genome sequencing technology becomes more accessible and affordable, much more efforts involving genome-wide sequencing of large number of tumor genomes will be conducted. Our work provides a conceptual and technical framework for incorporating the genome sequencing outputs and other types of data such as microarray profiles to get more insights into the cancer-signaling mechanisms that will facilitate the identification of key genes for biomarkers and drug development.

Materials and methods

Data sets used in this study

Human signaling network

To build up the human signaling network, we manually curated the signaling molecules (most of them are proteins) and the interactions between these molecules from the most comprehensive signaling pathway database, BioCarta (http://www.biocarta.com/). The pathways in the database are illustrated as diagrams. We manually recorded the names, functions, cellular locations, biochemical classifications and the regulatory (including activating and inhibitory) and interaction relations of the signaling molecules for each signaling pathway. To ensure the accuracy of the curation, all the data have been crosschecked four times by different researchers. After combining the curated information with another literature-mined signaling network that contains ∼500 signaling molecules (Ma'ayan *et al*, 2005), we obtained a signaling network containing ∼1100 proteins (Awan *et al*, 2007). We further extended this network by extracting and adding the signaling molecules and their relations from the Cancer Cell Map (http://cancer.cellmap.org/cellmap/), a database that contains 10 manually curated signaling pathways for cancer. As a result, the network contains 1634 nodes and 5089 links that include 2403 activation links (positive links), 741 inhibitory links (negative links), 1915 physical links (neutral links) and 30 links whose types are unknown (Supplementary Table 9). To our knowledge, this network is the biggest cellular signaling network at present.

Cancer mutated genes

The cancer mutated genes were taken from the COSMIC database (http://www.sanger.ac.uk/genetics/CGP/cosmic/) and other large-scale or genome-wide sequencing of tumor samples (Sjoblom *et al*, 2006; Greenman *et al*, 2007; Thomas *et al*, 2007). COSMIC database contains manually curated cancer mutated genes and the information of tumor samples, mutated sequences from literature and the output from the CGP's large-scale sequencing of tumor samples (Davies *et al*, 2005; Stephens *et al*, 2005; Greenman *et al*, 2007). The literature-curated genes were compiled as the Cancer Gene Census (Futreal *et al*, 2004), which is accessible in COSMIC database. The CGP is using human genome sequences and high-throughput mutation detection techniques to identify somatically acquired sequence mutations and hence to identify genes critical in the development of human cancers. A few recent publications (Davies *et al*, 2005; Stephens *et al*, 2005; Greenman *et al*, 2007) represent a small fraction of the CGP output. In addition, COSMIC database has provided mutation frequencies for most of the cancer mutated genes. The cancer gene mutation frequency of a gene is defined as the ratio of samples containing the mutated gene to the total samples screened for that gene. In the database, about one-third of the literature-curated mutated genes (Cancer Gene Census genes) have nonzero mutation frequencies, suggesting that the literature curation of these genes (i.e., included them into the Cancer Gene Census) has been supported by one or more other independent experiments.

For the network analysis in this study, we first intersected the network genes with the literature-curated mutated genes. As a result, we obtained 115 genes (Supplementary Table 10), of which 55 genes (Supplementary Table 10) have nonzero mutation frequencies. Meanwhile, we intersected the network genes with the mutated genes derived from the CGP large-scale sequencing output and several other genome-wide and high-throughput sequencing of tumor samples (Stephens *et al*, 2005; Sjoblom *et al*, 2006; Greenman *et al*, 2007; Thomas *et al*, 2007). As a result, we obtained another gene set containing 218 genes. Finally, we obtained 227 genes by merging the 55 genes and the 218 genes mentioned above. Among these 227 genes, 218 (96%) and 55 (24%) genes were collected from the large-scale sequencing of tumors and literature curation, respectively (Figure 1A). Notably, 46 genes (84%) of the literature-curated genes were overlapped with the mutated genes derived from the large-scale gene sequencing of tumors. The genes and their mutation frequencies from sequencing of tumors and literature were collected in Supplementary Table 10.

Methylated genes in cancer stem cells

We obtained 287 DNA-methylated genes from the three recent genome-wide determinations of the methylated genes from cancer stem cells (Ohm *et al*, 2007; Schlesinger *et al*, 2007; Widschwendter *et al*, 2007). Out of the 287 genes, 93 were mapped onto the human signaling network (Supplementary Table 11).

Cancer-associated gene set

The cancer-associated gene set contains the following data sources: (a) the cancer mutated genes we mentioned above; (b) a literature-mined breast cancer gene set from plasmID database (http://plasmid.hms.harvard.edu/GetCollectionList.do); (c) the genes extracted from the NCBI's Online Mendelian Inheritance in Man (OMIM) data set using the keywords such as 'cancer', 'tumor' and 'onco' etc. The cancer-associated gene list contains 2128 genes (Supplementary Table 8).

Microarray data

Gene expression profiles and the patients' survival data for the 295 breast tumor samples were obtained from Chang *et al* (2005).

Oncogenic map extraction

To extract an oncogenic map from the human signaling network, we mapped all the mutated and methylated genes onto the network. As a result, 67% of these genes are connected together to form a giant, linked network component. To include the mutated and methylated genes that are not present in this network component, we first found one shortest path between such a gene and a component node. If the length of the shortest path is 2 (i.e., the gene reaches one of the component nodes via a nonmutated network node), we linked that gene and the node on the shortest path into the component. A Java program had been written to implement this procedure (Supplementary File 1).

Network analysis

To extract the members of the branching and convergent units and 3-node- and 4-node-size network motifs, mfinder program (Kashtan *et al*, 2004) was used. To detect the signaling network communities from the oncogene-signaling map, we applied a network community algorithm (Newman, 2006).

Analyzing the enrichment of the mutated and methylated genes in the network motifs

We mapped the mutated and methylated genes onto each type of the motifs. We then counted the number of mutated or methylated genes in each motif and classified each type of motif into several subgroups based on the number of nodes that are mutated or methylated genes. We then calculated the ratio (Ra) of the activation links to the total activation and inhibitory links in each subgroup.

Randomization tests

We performed randomization tests to evaluate the statistical significance of the observations. A more detailed explanation of the randomization tests was described previously by Wang and Purisima (2005).

Survival analysis

To evaluate the prognostic value of a gene based on the gene expression profiles and the survival information of the tumor samples, we performed Kaplan–Meier analysis by implementing the Cox–Mantel log-rank test using R, a statistical computing language (http://www.r-project.org/). If the *P*-value is less than 0.05, the gene was thought as statistically significant to classify the tumor samples into 'good' and 'bad' groups.

Supplementary information

Supplementary information is available at the *Molecular Systems Biology* website (www.nature.com/msb).

Acknowledgements

This work was partially supported by the Genome and Heath Initiative. We thank Drs Z Yu, A Lenferink and the members of our team for discussion and comments.

References

Awan A, Bari H, Yan F, Mokin S, Yang S, Chowdhury Q, Yu Z, Purisima EO, Wang E (2007) Regulatory network motifs and hotspots of cancer genes in a mammalian cellular signaling network. *IET Syst Biol* **1:** 292–297

Babu MM, Luscombe NM, Aravind L, Gerstein M, Teichmann SA (2004) Structure and evolution of transcriptional regulatory networks. *Curr Opin Struct Biol* **14:** 283–291

Chang HY, Nuyten DS, Sneddon JB, Hastie T, Tibshirani R, Sorlie T, Dai H, He YD, van't Veer LJ, Bartelink H, van de RM, Brown PO, van d V (2005) Robustness, scalability, and integration of a wound-response gene expression signature in predicting breast cancer survival. *Proc Natl Acad Sci USA* **102:** 3738–3743

Chanock SJ, Thomas G (2007) The devil is in the DNA. *Nat Genet* **39:** 283–284

Chng WJ (2007) Limits to the human cancer genome project? *Science* **315:** 762–765

Cui Q, Yu Z, Purisima EO, Wang E (2006) Principles of microRNA regulation of a human cellular signaling network. *Mol Syst Biol* **2:** 46

Davies H, Hunter C, Smith R, Stephens P, Greenman C, Bignell G, Teague J, Butler A, Edkins S, Stevens C, Parker A, O'Meara S, Avis T, Barthorpe S, Brackenbury L, Buck G, Clements J, Cole J, Dicks E, Edwards K *et al* (2005) Somatic mutations of the protein kinase gene family in human lung cancer. *Cancer Res* **65:** 7591–7595

Ferrell JE (2002) Self-perpetuating states in signal transduction: positive feedback, double-negative feedback and bistability. *Curr Opin Cell Biol* **14:** 140–148

Futreal PA, Coin L, Marshall M, Down T, Hubbard T, Wooster R, Rahman N, Stratton MR (2004) A census of human cancer genes. *Nat Rev Cancer* **4:** 177–183

Greenman C, Stephens P, Smith R, Dalgliesh GL, Hunter C, Bignell G, Davies H, Teague J, Butler A, Stevens C, Edkins S, O'Meara S, Vastrik I, Schmidt EE, Avis T, Barthorpe S, Bhamra G, Buck G, Choudhury B, Clements J *et al* (2007) Patterns of somatic mutation in human cancer genomes. *Nature* **446:** 153–158

Gymnopoulos M, Elsliger MA, Vogt PK (2007) Rare cancer-specific mutations in PIK3CA show gain of function. *Proc Natl Acad Sci USA* **104:** 5569–5574

Han JD, Bertin N, Hao T, Goldberg DS, Berriz GF, Zhang LV, Dupuy D, Walhout AJ, Cusick ME, Roth FP, Vidal M (2004) Evidence for dynamically organized modularity in the yeast protein–protein interaction network. *Nature* **430:** 88–93

Hanahan D, Weinberg RA (2000) The hallmarks of cancer. *Cell* **100:** 57–70

Ihmels J, Levy R, Barkai N (2004) Principles of transcriptional control in the metabolic network of *Saccharomyces cerevisiae*. *Nat Biotechnol* **22:** 86–92

Ikediobi ON, Davies H, Bignell G, Edkins S, Stevens C, O'Meara S, Santarius T, Avis T, Barthorpe S, Brackenbury L, Buck G, Butler A, Clements J, Cole J, Dicks E, Forbes S, Gray K, Halliday K, Harrison R, Hills K *et al* (2006) Mutation analysis of 24 known cancer genes in the NCI-60 cell line set. *Mol Cancer Ther* **5:** 2606–2612

Jackson JG, Yoneda T, Clark GM, Yee D (2000) Elevated levels of p66 Shc are found in breast cancer cell lines and primary tumors with high metastatic potential. *Clin Cancer Res* **6:** 1135–1139

Kaiser J (2006) Cancer. First pass at cancer genome reveals complex landscape. *Science* **313:** 1370

Kashtan N, Itzkovitz S, Milo R, Alon U (2004) Efficient sampling algorithm for estimating subgraph concentrations and detecting network motifs. *Bioinformatics* **20:** 1746–1758

Kharchenko P, Church GM, Vitkup D (2005) Expression dynamics of a cellular metabolic network. *Mol Syst Biol* **1:** 2005

Liu ET, Lemberger T (2007) Higher order structure in the cancer transcriptome and systems medicine. *Mol Syst Biol* **3:** 94

Luscombe NM, Babu MM, Yu H, Snyder M, Teichmann SA, Gerstein M (2004) Genomic analysis of regulatory network dynamics reveals large topological changes. *Nature* **431:** 308–312

Ma'ayan A, Jenkins SL, Neves S, Hasseldine A, Grace E, Dubin-Thaler B, Eungdamrong NJ, Weng G, Ram PT, Rice JJ, Kershenbaum A, Stolovitzky GA, Blitzer RD, Iyengar R (2005) Formation of regulatory patterns during signal propagation in a mammalian cellular network. *Science* **309:** 1078–1083

Meuwissen R, Berns A (2005) Mouse models for human lung cancer. *Genes Dev* **19:** 643–664

Natarajan M, Lin KM, Hsueh RC, Sternweis PC, Ranganathan R (2006) A global analysis of cross-talk in a mammalian cellular signalling network. *Nat Cell Biol* **8:** 571–580

Newman ME (2006) Modularity and community structure in networks. *Proc Natl Acad Sci USA* **103:** 8577–8582

Nowell PC (1976) The clonal evolution of tumor cell populations. *Science* **194:** 23–28

Obata K, Morland SJ, Watson RH, Hitchcock A, Chenevix-Trench G, Thomas EJ, Campbell IG (1998) Frequent PTEN/MMAC mutations in endometrioid but not serous or mucinous epithelial ovarian tumors. *Cancer Res* **58:** 2095–2097

Ohm JE, McGarvey KM, Yu X, Cheng L, Schuebel KE, Cope L, Mohammad HP, Chen W, Daniel VC, Yu W, Berman DM, Jenuwein T, Pruitt K, Sharkis SJ, Watkins DN, Herman JG, Baylin SB (2007) A stem cell-like chromatin pattern may predispose tumor suppressor genes to DNA hypermethylation and heritable silencing. *Nat Genet* **39:** 237–242

Schlesinger Y, Straussman R, Keshet I, Farkash S, Hecht M, Zimmerman J, Eden E, Yakhini Z, Ben-Shushan E, Reubinoff BE, Bergman Y, Simon I, Cedar H (2007) Polycomb-mediated methylation on Lys27 of histone H3 pre-marks genes for *de novo* methylation in cancer. *Nat Genet* **39:** 232–236

Siegel PM, Massague J (2003) Cytostatic and apoptotic actions of TGF-beta in homeostasis and cancer. *Nat Rev Cancer* **3:** 807–821

Sjoblom T, Jones S, Wood LD, Parsons DW, Lin J, Barber TD, Mandelker D, Leary RJ, Ptak J, Silliman N, Szabo S, Buckhaults P, Farrell C, Meeh P, Markowitz SD, Willis J, Dawson D, Willson JK, Gazdar AF, Hartigan J *et al* (2006) The consensus coding sequences of human breast and colorectal cancers. *Science* **314:** 268–274

Stephens P, Edkins S, Davies H, Greenman C, Cox C, Hunter C, Bignell G, Teague J, Smith R, Stevens C, O'Meara S, Parker A, Tarpey P, Avis T, Barthorpe A, Brackenbury L, Buck G, Butler A, Clements J, Cole J *et al* (2005) A screen of the complete protein kinase gene family identifies diverse patterns of somatic mutations in human breast cancer. *Nat Genet* **37:** 590–592

Thomas RK, Baker AC, Debiasi RM, Winckler W, Laframboise T, Lin WM, Wang M, Feng W, Zander T, Macconaill LE, Lee JC, Nicoletti R, Hatton C, Goyette M, Girard L, Majmudar K, Ziaugra L, Wong KK, Gabriel S, Beroukhim R *et al* (2007) High-throughput oncogene mutation profiling in human cancer. *Nat Genet* **39:** 347–351

Wang E, Purisima E (2005) Network motifs are enriched with transcription factors whose transcripts have short half-lives. *Trends Genet* **21:** 492–495

Widschwendter M, Fiegl H, Egle D, Mueller-Holzner E, Spizzo G, Marth C, Weisenberger DJ, Campan M, Young J, Jacobs I, Laird PW (2007) Epigenetic stem cell signature in cancer. *Nat Genet* **39:** 157–158

Wuchty S, Oltvai ZN, Barabasi AL (2003) Evolutionary conservation of motif constituents in the yeast protein interaction network. *Nat Genet* **35:** 176–179

Zhang LV, King OD, Wong SL, Goldberg DS, Tong AH, Lesage G, Andrews B, Bussey H, Boone C, Roth FP (2005) Motifs, themes and thematic maps of an integrated *Saccharomyces cerevisiae* interaction network. *J Biol* **4:** 6