A multidimensional analysis of genes mutated in breast and colorectal cancers

Jimmy Lin, Christine M. Gan, Xiaosong Zhang, Siân Jones, Tobias Sjöblom, Laura D. Wood, D. Williams Parsons, Nickolas Papadopoulos, Kenneth W. Kinzler, Bert Vogelstein, Giovanni Parmigiani, and Victor E. Velculescu[1]

Ludwig Center for Cancer Genetics and Therapeutics, and The Howard Hughes Medical Institute at The Johns Hopkins Kimmel Cancer Center, Baltimore, Maryland 21231, USA

A recent study of a large number of genes in a panel of breast and colorectal cancers identified somatic mutations in 1149 genes. To identify potential biological processes affected by these genes, we examined their putative roles based on sequence similarity, membership in known functional groups and pathways, and predicted interactions with other proteins. These analyses identified functional groups and pathways that were enriched for mutated genes in both tumor types. Additionally, the results pointed to differences in molecular mechanisms that underlie breast and colorectal cancers, including various intracellular signaling and metabolic pathways. These studies provide a multidimensional framework to guide further research and help identify cellular processes critical for malignant progression and therapeutic intervention.

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

Cancer arises through the gradual accumulation of alterations in oncogenes and tumor suppressor genes. In an effort to identify such genes on a genomic scale, we have recently performed a systematic sequencing study of the majority of human genes in breast and colorectal cancers (Sjöblom et al. 2006). Analysis of 13,023 genes in 11 samples of each tumor type identified 1307 somatic (i.e., tumor-specific) mutations in 1149 genes. Using a statistical model that incorporated the mutation type, frequency, and sequence context, we identified a set of nearly 200 candidate cancer genes (*CAN*-genes) that were likely to play a driving role in tumorigenesis. In addition to the *CAN*-genes, there were additional mutated genes that may have been selected for during tumorigenesis, but which were mutated at a frequency that would not allow them to be distinguished from unselected passenger changes. The genes mutated in breast cancers were quite different from those mutated in colorectal cancers. Moreover, there were substantial differences in the mutated gene complement among any two samples of the same tumor type. Overall, this effort has identified a plethora of novel genes that are likely to play a role in human cancer. However, the study also suggested a higher level of complexity, in terms of both the number and type of genes involved, than previously thought to underlie the tumorigenic process.

Given this complexity, a systems biological approach could be useful to identify patterns among the mutated genes and to help interpret the genetic landscape of the two tumor types. An optimal approach of this sort would not only examine the individual roles of the mutated gene products, but would also explore their relationships, interactions, and network properties. Understanding this interplay could provide insight into mechanisms of tumorigenesis and prioritize specific pathways and processes for future genetic and biochemical research.

In this study, we take advantage of existing genomic and proteomic databases to highlight different aspects of the genes that are mutated in breast and colorectal cancers. Our analysis uses four different system-level perspectives: (1) sequence similarity, (2) functional annotation (including cellular function, biochemical processes, and subcellular localization), (3) protein–protein interactions, and (4) molecular pathways. At each of these levels, we identify specific gene groups that were enriched for genetic alterations, revealing potentially aberrant cellular processes in the tumors.

[1]Corresponding author.
E-mail velculescu@jhmi.edu; fax (410) 955-0548.
Article published online before print. Article and publication date are at http://www.genome.org/cgi/doi/10.1101/gr.6431107.

Results

Protein sequence similarity

We first evaluated the proteins encoded by the 1149 mutated genes through sequence-similarity analyses. This approach provides an unbiased means to group proteins based on their encoded information content. Two complementary methods were used: pairwise basic local alignment search tool (BLAST) analysis and comparison of protein domains using information from existing databases. Sequence comparisons via BLAST facilitated examination of entire coding regions, while analyses of protein domains identified motifs and sequence relationships that would not be evident through whole gene comparisons.

To compare entire coding regions we used BLASTP (Altschul et al. 1990) to compare sequences of all mutant proteins and constructed protein networks based on high-sequence similarity (Fig. 1). These networks identified clusters of proteins, each labeled according to the predominant functional role of the proteins contained. From a global perspective, breast and colorectal cancers shared many common clusters, including zinc finger proteins, cadherins, and genes involved in cell adhesion and signal transduction. Clusters that were mutated in one tumor type but not the other included semaphorins, RNA helicases, and DNA helicases in breast cancers and metalloproteinases, voltage-gated K+ channels, and orphan G-protein coupled receptors in colorectal cancers.

17:1304–1318 ©2007 by Cold Spring Harbor Laboratory Press; ISSN 1088-9051/07; www.genome.org

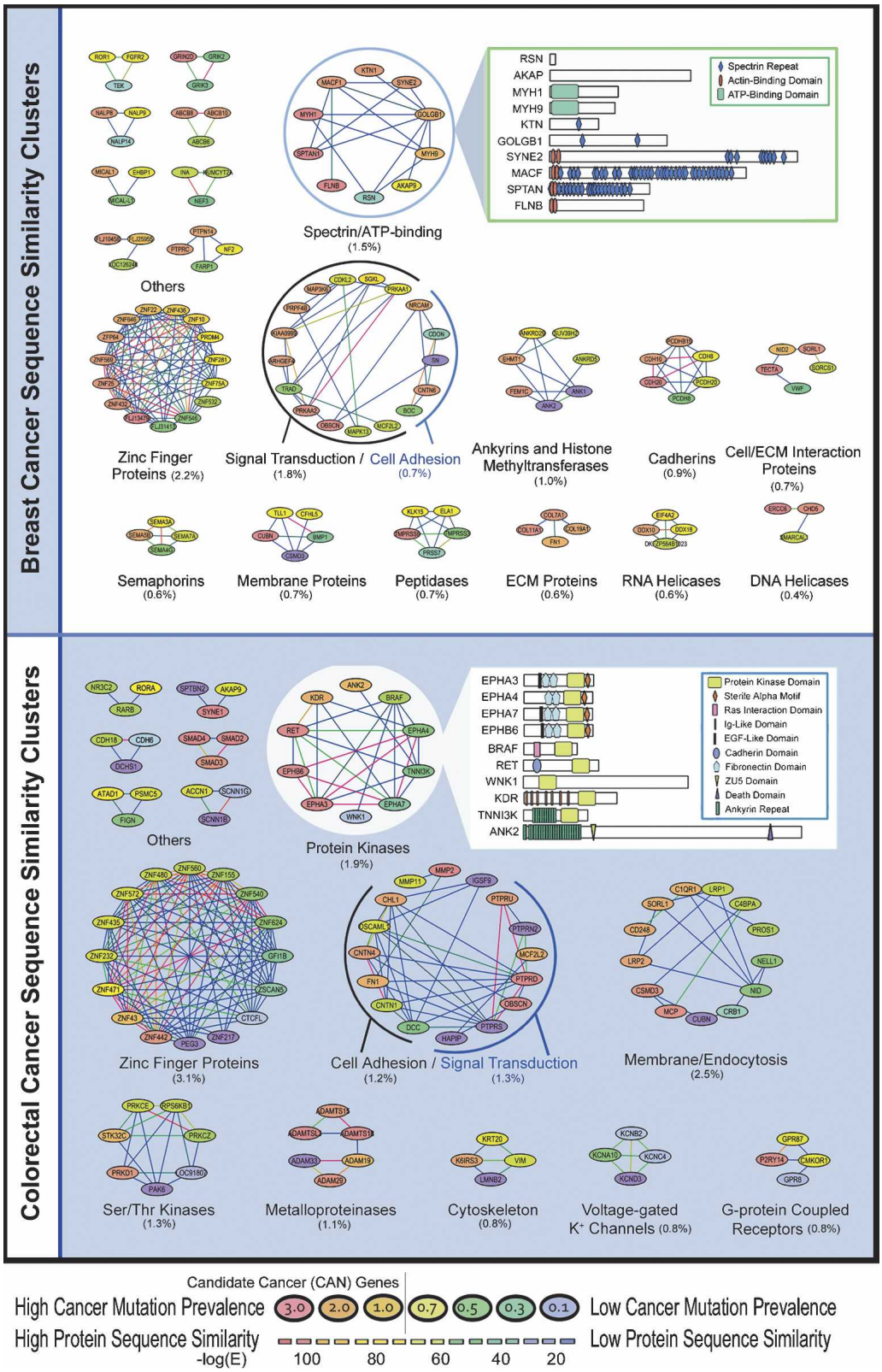

Figure 1. Sequence similarity among mutated genes in breast and colorectal cancers. Each cluster represents genes that are mutated in breast (*top*) or colorectal cancers (*bottom*). Each node represents a gene that is colored according to the Cancer Mutation Prevalence Score (CaMP score), and each line represents a sequence-similarity relationship that is colored according to degree of sequence similarity. *CAN*-genes identified by Sjöblom et al. (2006) have a CaMP score >1 and are colored in orange and red. Clusters are named according to the predominant genes contained within each cluster, and those containing only two genes are not shown. The percentage of the total mutated genes contained within each cluster is showed in parentheses. The inset highlights local similarity within protein domains of genes in a specific cluster.

Genes that have high sequence identity often participate in similar intracellular roles, either through related biochemical functions, protein dimerization, genetic interactions, or more complex relationships. Within the clusters shown in Figure 1 there were several instances of patterns suggesting common functions during tumorigenesis. For example, mutations in ephrin receptors *EPHA3*, *EPHA4*, *EPHA7*, or *EPHB6* affected 10 of the 35 colorectal tumors examined, but no tumor contained mutations in more than a single ephrin receptor, suggesting mutual exclusivity among mutations in these genes. Global analyses of sequence-similarity clusters in both breast and colorectal cancers identified nine and four clusters that showed mutual exclusivity, respectively. While the genes within some pathways act in series, and mutation of one member of the pathway is sufficient to disrupt function, clusters of sequence similarity may also include members that act in parallel pathways. For example, mutations in the TGF-beta pathway mediators *SMAD2*, *SMAD3*, or *SMAD4* occurred in seven of 35 tumors. While mutations in *SMAD4* did not occur in tumors with other *SMAD* mutations, both *SMAD2* and *SMAD3* were co-mutated in colorectal tumors Mx30 and Hx5 (Supplemental Fig. 1). Interestingly, SMAD2 or SMAD3 can separately heterodimerize with SMAD4 transcription factors upon pathway activation and mediate transcriptional responses (Jayaraman and Massague 2000). These results suggest that inactivation of either *SMAD4* alone or *SMAD2* and *SMAD3* together have similar effects on the TGF-beta receptor pathway.

A complementary method for analysis of sequence similarity takes advantage of information from existing databases. Instead of determining relatedness solely using BLASTP, other methods such as Hidden Markov Models and consensus sequences have facilitated in-depth comparisons of protein sequences. The Integrated Resource of Protein Families, Domains, and Sites (InterPro) database incorporates information from 16 protein databases, including Pfam, ProDom, PRINTS, PROSITE, and SMART (Apweiler et al. 2001). Using the annotation provided by InterPro 13.0, we examined the protein sequences of all mutated genes for the presence of specific domains. A total of 13,147 possible domains were examined in 1149 mutated proteins, resulting in a total of 1029 proteins that were found to have 3549 domain assignments.

We examined these data in two ways to determine whether gene groups containing specific domains were more likely to be mutated than predicted by chance alone. First, we determined whether the number of mutations in gene groups containing specific domains reflected a mutation prevalence that was significantly higher than the passenger mutation prevalence. We performed these calculations for breast and colorectal cancers separately, using the conservative assumption that the observed mutation frequencies of 2.5 and 3.3 mutations per million base pairs, respectively, constituted the passenger rates. Note that this criterion is highly conservative, as the observed mutations actually represent the sum of passenger mutations and those mutations selected for during tumorigenesis (i.e., pathogenic mutations). The resulting Group CaMP score is similar to that used to derive the Cancer Mutation Prevalence (CaMP) score for individual genes. The Group CaMP score incorporated the total number of mutations from all genes within each group, the combined lengths of the genes in each group, and the total number of tumors examined. The *P*-value of observing at least the number of mutations in a binomial distribution was calculated and corrected with the Benjamini–Hochberg algorithm (Benjamini and Hochberg 1995).

Second, we examined whether the distributions of individual CaMP scores of mutated genes containing domains of interest were different from mutated genes not containing such domains. To compare such distributions, we adapted the Gene Set Enrichment Analysis (GSEA) algorithm, using CaMP scores of individual genes instead of summaries of gene-expression values (Subramanian et al. 2005). The CaMP GSEA approach incorporates the set of individual CaMP scores from all genes within each group, accounting for the number, type and context of mutations observed, gene length, as well as the total number of tumors examined. This approach is complementary to the group CaMP score described above; while the former approach is more sensitive to the overall mutation prevalence in a group of genes, the latter would be expected to identify more subtle differences among the mutated genes within such groups.

After identification of candidate groups that were significantly enriched for mutations using these approaches, we filtered the results to identify those groups that were also enriched for an increased number of mutant genes. Specifically, we determined whether the ratio of the number of mutant genes containing each specific domain to all genes containing that domain was statistically higher than the ratio of the total number of mutant genes (1149) to the number of all the genes (13,023) analyzed. This filtering step ensured that multiple genes within each gene group must be affected in order for the entire group to be considered of interest. A gene group that had contained only one highly mutated gene (e.g., mutations only in TP53) would thereby be excluded.

Using these two analysis approaches (Group CaMP and CaMP GSEA), a total of 31 and 22 InterPro domains were significantly associated with colorectal and breast cancers, respectively (Table 1; Supplemental Table 1). In colorectal cancers, the majority were determined to be significant by both methods and involved several related protein domains. For example, 14 of the identified domains are in proteins that have extracellular regions or are involved in cell–cell interactions (e.g., four immuloglobulin-related domains, two fibronectin domains, six EGF-related domains, and two cadherin-related domains). An additional five domains (e.g., pleckstrin-like domain, DH domain, Ephrin receptor ligand-binding domain, Sterile alpha motif homology 2, and receptor tyrosine kinase domain) are known to be involved in protein kinase or G protein signal transduction pathways. Domains identified that were associated with metalloproteases include reprolysin, peptidase M12B propeptide, cysteine-rich ADAM, and disintegrin. Finally, domains present in TGF-beta pathway transcription mediators SMAD (MAD homology 1 and MAD homology 2 domains) were also identified as significantly associated with colorectal cancer. Interestingly, proteins containing MAD homology, ephrin receptor, and Treacher Collins Syndrome protein domains were found to be exclusively mutated in colorectal cancers, while members of the other domains were mutated in both tumor types. Other domains shared by both cancer types include three of the extracellular EGF-related domains, as well as two domains involved in signaling, the DH domain and the pleckstrin-like domain. In breast cancers, two motifs were detected by both the Group CaMP and GSEA methods: one was the spectrin repeat domain that is present in various cytoskeletal proteins, while the second was the relatively nonspecific proline-rich region domain that was also associated with colorectal cancers. Three domains related to ABC transporters and two domains involved in actin binding were preferentially identified in breast tumors.

Table 1. Protein domains identified through mutation enrichment analysis

	Significance[a]		Protein domain		Mutation enrichment						Gene enrichment filter					
Tissue	Group CaMP	CaMP GSEA	InterPro accession	Domain description	No. of mutations	Base pairs sequenced (Mb)	Expected no. of mutations	Percentage of mutations within domain	Group CaMP score	CaMP GSEA score	No. of genes mutated	No. of genes sequenced	Expected no. of genes mutated	Fold change	χ^2 *P*-value	Composite score[b]
Colorectal	●	●	IPR000977	ATP-dependent DNA ligase	8	309,843	0.9	25%	3.34	>20	4	5	0.2	20.1	2.37×10^{-4}	468.5
	●	●	IPR001132	MAD homology 2, Dwarfin-type	9	105,902	0.3	44%	8.06	>20	3	7	0.3	10.8	5.57×10^{-3}	301.7
	●	●	IPR003619	MAD homology 1, Dwarfin-type	9	151,572	0.4	11%	6.76	>20	3	10	0.4	7.5	1.22×10^{-2}	201.5
	●	●	IPR000033	Low-density lipoprotein receptor, YWTD repeat	8	683,617	1.9	25%	1.19	>20	4	11	0.4	9.1	2.13×10^{-3}	193.3
	●	●	IPR003962	Fibronectin, type III subdomain	27	2,308,840	6.5	0%	6.21	>20	13	59	2.4	5.5	3.66×10^{-6}	144.9
	●	●	IPR002870	Peptidase M12B, propeptide	10	707,566	2.0	10%	2.29	>20	5	23	0.9	5.5	3.96×10^{-3}	121.6
	●	●	IPR001090	Ephrin receptor, ligand binding	10	358,351	1.0	20%	4.72	7.57	4	12	0.5	8.4	2.75×10^{-3}	102.8
	●	●	IPR001426	Receptor tyrosine kinase, class V	10	358,351	1.0	0%	4.72	7.57	4	12	0.5	8.4	2.75×10^{-3}	102.8
	●	●	IPR003961	Fibronectin, type III	38	3,911,114	11.0	32%	6.96	>20	17	112	4.5	3.8	1.14×10^{-5}	102.7
	●	●	IPR001590	Peptidase M12B, ADAM/reprolysin	10	832,715	2.3	40%	1.79	13.29	5	27	1.1	4.6	7.14×10^{-3}	70.1
		●	IPR006586	ADAM, cysteine-rich	5	300,840	0.8	0%	0.94	7.44	3	12	0.5	6.3	1.83×10^{-2}	52.6
		●	IPR001762	Disintegrin	5	363,720	1.0	0%	0.64	5.63	3	15	0.6	5.0	3.01×10^{-2}	31.5
	●	●	IPR013098	Immunoglobulin I-set	30	3,541,098	9.9	27%	4.16	5.32	15	117	4.7	3.2	1.96×10^{-4}	30.5
	●	●	IPR013032	EGF-like region	36	6,371,666	17.8	0%	1.69	9.22	21	196	7.8	2.7	1.29×10^{-4}	29.3
	●	●	IPR000219	DH	19	1,792,057	5.0	5%	3.52	3.33	8	47	1.9	4.3	1.18×10^{-3}	29.3
		●	IPR013091	EGF calcium binding	14	2,073,828	5.8	0%	0.66	5.63	9	54	2.2	4.2	6.81×10^{-4}	26.3
	●	●	IPR013106	Immunoglobulin V-set	35	4,432,992	12.4	6%	4.33	4.87	18	202	8.1	2.2	2.49×10^{-3}	20.6
	●		IPR003993	Treacher Collins syndrome protein Treacle	7	276,848	0.8	14%	2.73	0.48	4	19	0.8	5.3	1.08×10^{-2}	16.9
	●	●	IPR011510	Sterile alpha motif homology 2	14	1,483,790	4.2	7%	1.87	2.87	7	52	2.1	3.4	7.44×10^{-3}	16.0
		●	IPR001881	EGF-like calcium binding	14	2,501,013	7.0	7%	0.17	4.18	9	67	2.7	3.4	2.61×10^{-3}	14.6
		●	IPR000152	Aspartic acid and asparagine hydroxylation site	14	2,845,810	8.0	0%	−0.11	3.65	9	73	2.9	3.1	4.37×10^{-3}	11.0
	●	●	IPR000694	Proline-rich region	98	22,008,620	61.6	0%	2.06	3.58	54	925	36.9	1.5	7.67×10^{-3}	8.3
	●	●	IPR013151	Immunoglobulin	31	4,927,391	13.8	16%	2.02	1.69	16	219	8.7	1.8	2.12×10^{-2}	6.8
		●	IPR006209	EGF-like	18	4,353,587	12.2	0%	−0.33	2.99	12	121	4.8	2.5	5.60×10^{-3}	6.6
		●	IPR013111	EGF, extracellular	12	2,376,926	6.7	0%	−0.15	1.91	8	66	2.6	3.0	7.65×10^{-3}	5.3
		●	IPR000742	EGF-like, type 3	18	4,479,751	12.5	6%	−0.38	2.65	12	131	5.2	2.3	9.76×10^{-3}	5.2
	●		IPR001849	Pleckstrin-like	28	4,714,281	13.2	11%	1.41	0.66	16	162	6.5	2.5	1.63×10^{-3}	5.1
		●	IPR013585	Protocadherin	11	2,341,227	6.6	0%	−0.30	1.87	7	68	2.7	2.6	2.55×10^{-2}	4.0
		●	IPR002126	Cadherin	11	2,366,651	6.6	36%	−0.32	1.82	7	69	2.7	2.5	2.72×10^{-2}	3.8

(continued)

Table 1. *Continued*

	Significance[a]		Protein domain		Mutation enrichment						Gene enrichment filter					
Tissue	Group CaMP	CaMP GSEA	InterPro accession	Domain description	No. of mutations	Base pairs sequenced (Mb)	Expected no. of mutations	Percentage of mutations within domain	Group CaMP score	CaMP GSEA score	No. of genes mutated	No. of genes sequenced	Expected no. of genes mutated	Fold change	χ^2 *P*-value	Composite score[b]
Breast	●	●	IPR002017	Spectrin repeat	22	2,251,242	7.9	41%	1.15	>20	8	34	1.8	4.6	9.23×10^{-4}	96.3
		●	IPR000998	MAM	6	339,166	1.2	17%	−0.93	6.64	3	10	0.5	5.8	2.36×10^{-2}	33.1
	●	●	IPR000694	Proline-rich region	140	21,894,664	76.6	0%	9.16	7.30	78	925	47.8	1.6	1.18×10^{-4}	26.9
		●	IPR001849	Pleckstrin-like	38	4,687,506	16.4	8%	0.55	10.09	21	162	8.4	2.5	3.27×10^{-4}	26.7
		●	IPR001589	Actin binding, actinin-type	13	1,233,393	4.3	8%	−0.29	4.93	5	27	1.4	3.6	1.93×10^{-2}	16.6
		●	IPR002172	Low-density lipoprotein-receptor, class A	11	1,019,974	3.6	0%	−1.14	4.76	5	28	1.4	3.5	2.18×10^{-2}	12.5
		●	IPR005479	Carbamoyl-phosphate synthase L chain, ATP binding	6	619,398	2.2	0%	−3.57	5.76	3	11	0.6	5.3	2.89×10^{-2}	11.6
		●	IPR001715	Calponin-like actin binding	17	1,823,187	6.4	6%	−0.41	2.50	8	43	2.2	3.6	3.35×10^{-3}	7.5
		●	IPR003439	ABC transporter related	14	1,843,971	6.5	7%	−2.16	4.42	8	49	2.5	3.2	6.69×10^{-3}	7.1
		●	IPR001164	Arf GTPase activating protein	8	531,103	1.9	0%	−1.23	2.37	4	22	1.1	3.5	3.70×10^{-2}	4.0
		●	IPR000533	Tropomyosin	22	3,410,366	11.9	5%	−1.10	1.96	15	118	6.1	2.5	2.52×10^{-3}	2.1
		●	IPR001140	ABC transporter, transmembrane region	5	686,959	2.4	20%	−3.15	2.92	3	19	1.0	3.1	9.15×10^{-2}	−0.7
		●	IPR000225	Armadillo	6	637,473	2.2	17%	−2.33	1.80	4	26	1.3	3.0	5.81×10^{-2}	−1.6
		●	IPR013563	Oligopeptide/dipeptide ABC transporter, C-terminal	9	1,454,816	5.1	0%	−3.82	3.07	5	38	2.0	2.5	5.92×10^{-2}	−1.9
		●	IPR000357	HEAT	10	1,776,863	6.2	0%	−2.53	1.80	7	52	2.7	2.6	2.57×10^{-2}	−1.9
		●	IPR000219	DH	16	1,779,811	6.2	13%	−2.75	2.18	10	47	2.4	4.1	4.43×10^{-4}	−2.4
		●	IPR003961	Fibronectin, type III	26	3,900,508	13.7	27%	−4.93	3.66	14	112	5.8	2.4	3.93×10^{-3}	−3.1
		●	IPR000909	Phosphatidylinositol-specific phospholipase C, X region	4	276,847	1.0	25%	−2.13	1.34	3	10	0.5	5.8	2.36×10^{-2}	−4.6
		●	IPR013111	EGF, extracellular	16	2,368,216	8.3	0%	−3.23	1.15	8	66	3.4	2.3	2.93×10^{-2}	−4.9
		●	IPR013032	EGF-like region	31	6,336,135	22.2	3%	−5.40	1.15	17	196	10.1	1.7	3.63×10^{-2}	−7.1
		●	IPR013091	EGF calcium binding	12	2,064,023	7.2	0%	−5.23	1.72	6	54	2.8	2.2	7.46×10^{-2}	−7.5
		●	IPR003962	Fibronectin, type III subdomain	13	2,301,425	8.1	0%	−7.49	3.33	8	59	3.0	2.6	1.71×10^{-2}	−10.9

[a]Domains wth Group CaMP scores or CaMP GSEA scores >1 were considered significant. The false discovery rates for each of these approaches is estimated to be 10% (see Methods for details).
[b]The composite score was determined to be the product of the sum of the Group CaMP and CaMP GSEA scores and the fold change of the gene enrichment.

Functional annotation and gene ontology

In addition to analyses based on sequence content, the mutated genes were categorized according to their annotated biological roles. The Gene Ontology (GO) Consortium has devised a controlled vocabulary for describing molecular functions and biological processes of genes based on information obtained from the literature and from sequence and biological databases (Ashburner et al. 2000). These are represented in hierarchical levels of directed acyclic relationships that progress from general descriptions to progressively more specific descriptions. In general, a gene can have multiple descriptions, and functional descriptions for any gene using GO can be complex. To simplify such descriptions, methods have been designed to summarize the GO relationships into fewer general categories (Camon et al. 2004; Martin et al. 2004). For these analyses, we first used a general approach to examine the broad functional categories of all mutated genes in breast and colorectal cancers (Fig. 2), and then identified the specific GO groups that were preferentially associated with each of these tumor types (Table 2).

Classification of the mutated genes into general functional categories was visualized using OSPREY (Fig. 2) (Breitkreutz et al. 2003). In accord with our initial analysis of a small subset of mutated genes (Sjöblom et al. 2006), the comprehensive analyses of all mutated genes resulted in similar compositions of functional categories for both breast and colorectal cancers. The fractions of different functional categories were largely comparable between the two cancer types, with the two largest comprising signal transduction and metabolism. Importantly, the individual genes that were included in these categories were different in the different tumor types, and individual tumors had a varying composition of genes belonging to these functional categories (Supplemental Fig. 2). Noticeably, over a third of the mutant genes were not assigned to any functional category using the current annotation, a fraction that would be expected to decrease as additional biological data are obtained. A similar analysis of the subset of genes most highly mutated in breast and colorectal cancers identified subtle differences in composition of functional categories (Fig. 2). These analyses suggest that genes that are selected for mutation in human cancer may come from a variety of different functional categories, but that such broad categories may not be helpful in accurately capturing specific functional aspects of genes that are preferentially mutated in tumors.

In order to identify more specific molecular functions for the mutant genes, we examined the full set of 18,740 GO groups. Using approaches similar to that used in the analysis of protein domains, we identified GO groups that were enriched for the number of mutations or distribution of CaMP scores using CaMP GSEA and Group CaMP approaches. In colorectal cancer, we identified 11 GO groups to be significant by either method (Table 2; Supplemental Table 2). Groups such as ephrin receptor activity as well as metalloendopeptidase activity corroborated results identified above through the analysis of protein domains. Two of the largest functional groups, cell adhesion and receptor activity, had 24 and 39 mutated genes and 60 and 63 mutations, respectively. More specific subgroups from these groups included insulin receptor binding and homophilic cell adhesion.

In breast cancers, 15 functional groups were identified, none overlapping precisely with those of colon. The most closely related ones involved functional groups that were involved in cell adhesion. The largest group identified was calcium ion binding, which included 50 mutated genes and 77 mutations. Five groups were associated with the extracellular matrix, including extracellular matrix organization and biogenesis, extracellular matrix structural constituent, microtubule binding, actin binding, and cell–cell adhesion. Interestingly, two metabolic groups were affected in breast tumors: the overlapping groups of the urea cycle and arginine biosynthesis. Finally, there were three groups related to G protein signaling: GTPase activator activity and two Rho protein modulating groups. These analyses clearly show that while overall functional patterns may be similar between breast and colorectal cancers, the specific group constituents of these general categories are quite different.

Protein interactions

As part of their biologic roles, many proteins physically interact with other proteins; such interactions can form protein complexes or may represent functional components of molecular pathways. An additional perspective on gene function can therefore be realized through the analysis of physical interactions. Efforts to identify interacting proteins at a genome-wide scale, termed interactomes, have been performed for yeast (Uetz et al. 2000; Ito et al. 2001), fly (Giot et al. 2003), and worm (Li et al. 2004). Although the human interactome is incomplete, human protein–protein interactions can be inferred from data on model organisms using interologs (human protein orthologs that are known to interact in other organisms). Such interologs have differing levels of reliability, in part based on the strength of the interaction, the number of organisms in which the interaction is observed, and the sequence similarity of the interacting proteins to their human counterparts. Using this approach, we examined the predicted interactions of mutated proteins with other proteins. Overall, this analysis showed that the mutant proteins interacted with more partners than typical human proteins. Proteins encoded by CCDS genes had an average of 11.8 predicted interactions each, while the set of the 189 most highly mutated *CAN*-genes had an average of 26.0 and 16.4 interactions in colon and breast cancers, respectively ($P = 0.02$ and 0.12, Student's *t*-test). Proteins with a large number of interactions have been suggested to serve as essential hubs of molecular pathways such as those that are disrupted in cancers (Jonsson and Bates 2006).

To identify networks of interacting proteins that were preferentially altered in cancers, we analyzed the predicted interactions of mutated proteins in each tumor type (Fig. 3; Supplemental Figs. 3, 4). In breast cancers, over half of the mutated proteins (59 of 83) were predicted to participate in a large interaction cluster driven by links to TP53, BRCA1, PIK3R1, and NFKB. In contrast, the largest interaction cluster in colorectal cancers involved SMAD proteins and contained only 12 proteins, and the only cluster containing more than five proteins included TP53. These analyses emphasize how mutation studies coupled with systems analysis can provide information useful for understanding the pathways through which the mutant proteins function. For example, the mutation interactome highlighted three interacting SMAD proteins in colorectal cancers and a cluster of circadian rhythm proteins (PER1, PER2, and TIMELESS) in breast cancers. The proteins encoded by the latter three genes are thought to control cell cycle progression, and genetic inactivation of one of the genes (*Per2*) has been shown to lead to tumor predisposition in mice (Fu et al. 2002; Chen et al. 2005; Lee 2006).

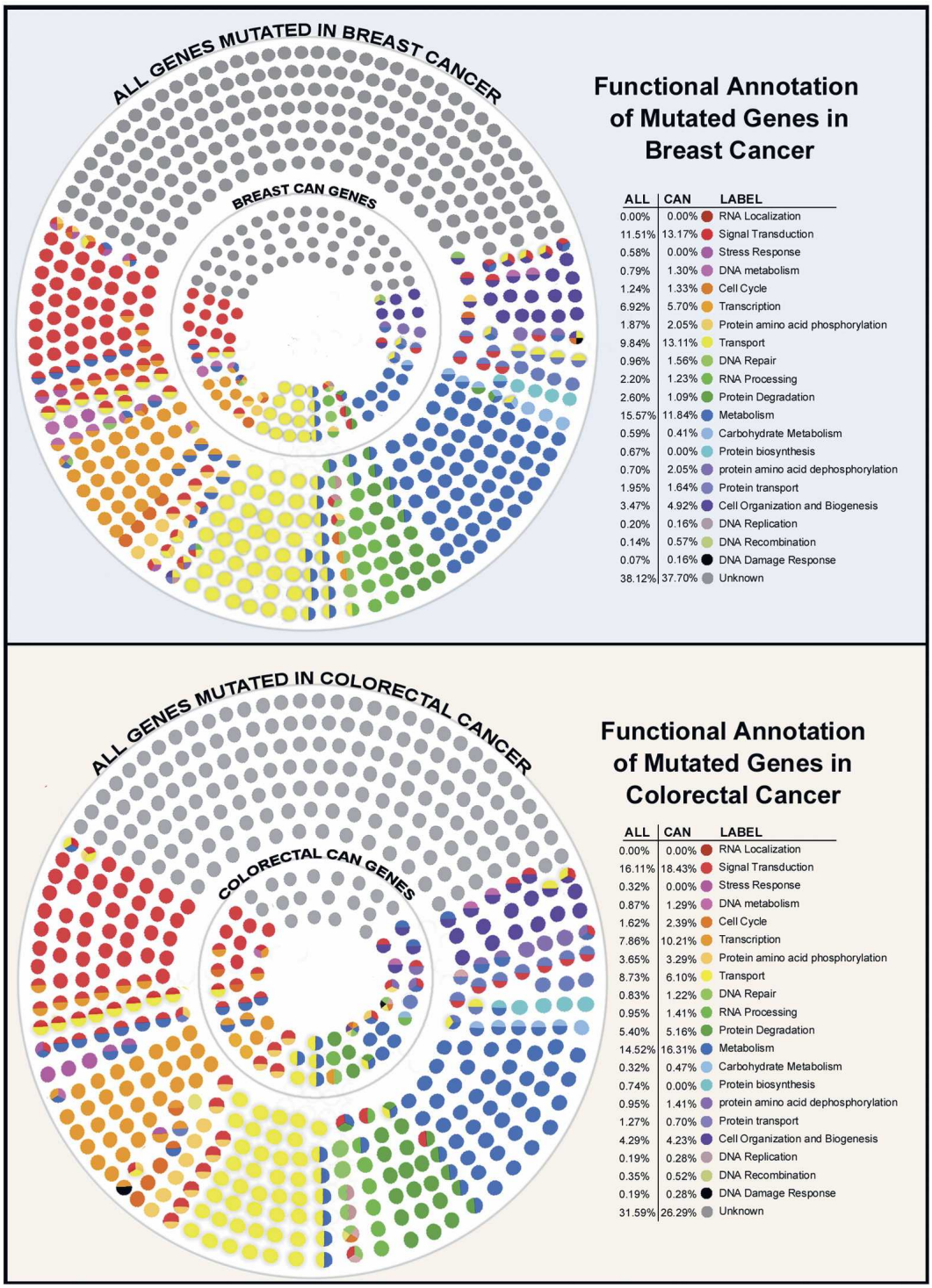

Figure 2. General functional categorization of genes mutated in breast and colorectal cancers. Each small circle represents a mutated gene in breast or colorectal cancer and is colored according to the general functional categories shown in the legend (for details, see Methods). The entire set of circles represents all the genes mutated in each cancer type, while the interior subset is comprised of the genes with the highest CaMP Scores (the *CAN*-genes). The percentage of genes that belong to each functional category is shown in the legend.

Molecular pathways

Pathways can be defined as the stepwise interaction of multiple proteins designed to achieve a defined cellular process. A variety of signaling, metabolic, and other pathways have been cataloged by the Kyoto Encylopedia of Genes and Genomes (KEGG) (Ogata et al. 1999), the iPATH (http://escience.invitrogen.com/ipath/), BioCarta (http://www.biocarta.com/), and sigPathways (Tian et

Table 2. Gene Ontology groups identified through mutation enrichment analysis

	Significance[a]		Gene Ontology groups		Mutation enrichment					Gene enrichment filter					
Tissue	Group CaMP	CaMP GSEA	GO ID	GO description	No. of mutations	Base pairs sequenced (Mb)	Expected no. of mutations	Group CaMP score	CaMP GSEA score	No. of genes mutated	No. of genes sequenced	Expected no. of genes mutated	Fold change	χ^2 *P*-value	Composite score[b]
Colorectal	●	●	GO:0005003	Ephrin receptor activity	10	846,787	2.4	3.00	10.44	4	9	0.4	11.2	1.19×10^{-3}	149.9
		●	GO:0051018	Protein kinase A binding	5	930,542	2.6	−0.51	11.65	3	9	0.4	8.4	9.63×10^{-3}	93.2
		●	GO:0005158	Insulin receptor binding	5	814,541	2.3	−0.06	6.59	3	14	0.6	5.4	2.58×10^{-2}	35.1
	●		GO:0045786	Negative regulation of progression through cell cycle	37	3,604,199	10.1	9.30	−0.82	7	74	2.9	2.4	3.67×10^{-2}	20.1
		●	GO:0004222	Metalloendopeptidase activity	11	3,521,487	9.9	−0.42	5.90	6	59	2.4	2.6	3.89×10^{-2}	14.0
		●	GO:0006898	Receptor mediated endocytosis	6	2,003,081	5.6	−0.92	3.30	3	24	1.0	3.1	8.37×10^{-2}	7.4
		●	GO:0007155	Cell adhesion	60	17,368,246	48.6	0.34	2.23	24	300	12.0	2.0	2.13×10^{-2}	5.2
		●	GO:0008203	Cholesterol metabolism	5	2,248,581	6.3	−1.09	2.71	3	26	1.0	2.9	9.89×10^{-2}	4.7
		●	GO:0007156	Homophilic cell adhesion	11	5,680,191	15.9	−1.02	1.31	7	79	3.1	2.2	4.82×10^{-2}	0.6
		●	GO:0008202	Steroid metabolism	6	2,828,563	7.9	−1.04	1.31	4	42	1.7	2.4	9.98×10^{-2}	0.6
		●	GO:0004872	Receptor activity	63	30,821,371	86.3	−0.84	1.31	39	751	29.9	1.3	7.50×10^{-2}	0.6
Breast		●	GO:0030198	Extracellular matrix organization and biogenesis	8	879,358	3.1	0.88	>20	3	11	0.6	5.3	2.89×10^{-2}	110.2
		●	GO:0006526	Arginine biosynthesis	4	432,452	1.5	−0.13	2.32	3	6	0.3	9.7	8.03×10^{-3}	21.2
		●	GO:0042626	ATPase activity, coupled to transmembrane movement	9	2,306,247	8.1	−0.54	5.88	5	25	1.3	3.9	1.48×10^{-2}	20.7
		●	GO:0000050	Urea cycle	4	339,644	1.2	0.64	1.50	3	8	0.4	7.3	1.46×10^{-2}	15.5
		●	GO:0016337	Cell–cell adhesion	10	1,873,173	6.6	−0.55	4.83	5	27	1.4	3.6	1.93×10^{-2}	15.3
		●	GO:0004722	Protein serine/threonine phosphatase activity	6	792,458	2.8	0.20	2.99	3	18	0.9	3.2	8.19×10^{-2}	10.3
		●	GO:0016887	ATPase activity	17	7,122,833	24.9	−0.93	5.15	8	71	3.7	2.2	4.09×10^{-2}	9.2
		●	GO:0003779	Actin binding	29	8,414,192	29.4	−0.64	4.80	13	128	6.6	2.0	2.29×10^{-2}	8.2
		●	GO:0005201	Extracellular matrix structural constituent	13	3,663,856	12.8	−0.62	3.98	6	53	2.7	2.2	6.99×10^{-2}	7.4
		●	GO:0005096	GTPase activator activity	14	3,744,913	13.1	−0.51	2.57	10	69	3.6	2.8	5.50×10^{-3}	5.8
		●	GO:0035023	Regulation of Rho protein signal transduction	12	4,099,034	14.3	−1.07	2.24	9	50	2.6	3.5	2.33×10^{-3}	4.1
		●	GO:0005089	Rho guanyl-nucleotide exchange factor activity	12	4,099,034	14.3	−1.07	2.24	9	50	2.6	3.5	2.33×10^{-3}	4.1
		●	GO:0005509	Calcium ion binding	77	32,098,919	112.3	−0.85	2.42	50	590	30.5	1.6	1.37×10^{-3}	2.6
		●	GO:0008017	Microtubule binding	9	3,061,703	10.7	−2.33	1.80	5	26	1.3	3.7	1.69×10^{-2}	−2.0

[a]GO Groups wth Group CaMP scores or CaMP GSEA scores >1 were considered significant. The false discovery rates for each of these approaches is estimated to be 10% (see Methods for details).
[b]The composite score was determined to be the product of the sum of the Group CaMP and CaMP GSEA scores and the fold change of the gene enrichment.

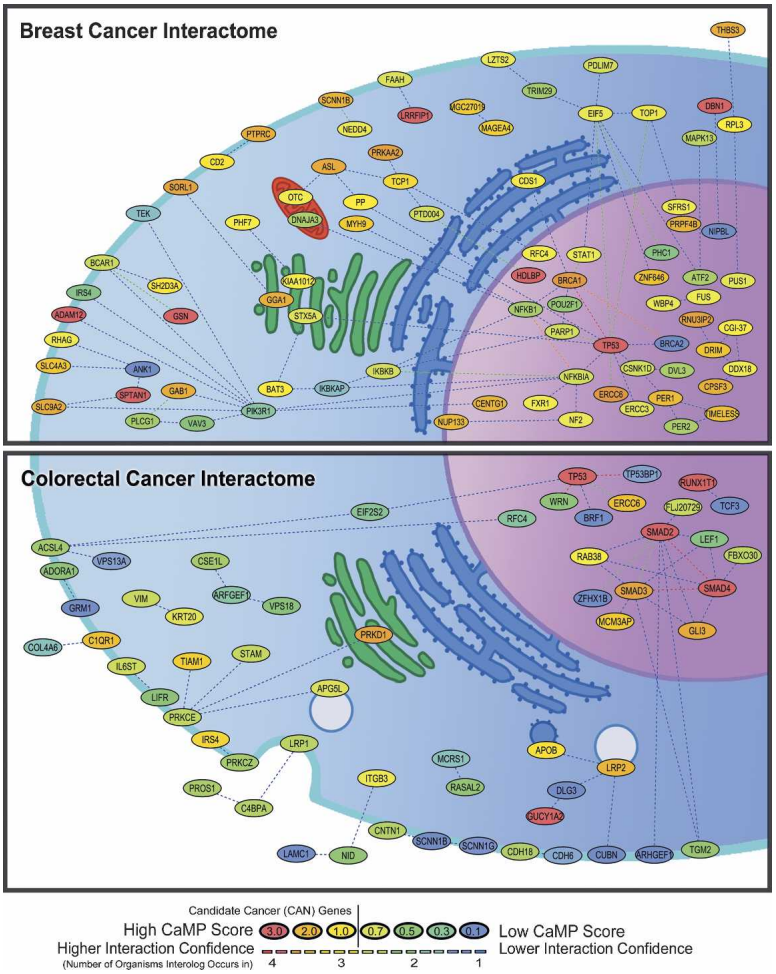

Figure 3. Interaction among proteins mutated in breast and colorectal cancers. Each node represents a mutated protein that is colored according to Cancer Mutation Prevalence (CaMP) Score, and each line represents an interaction confidence. *CAN*-genes identified by Sjöblom et al. (2006) have a CaMP score >1 and are colored in orange and red. The genes are placed within cellular compartments as annotated in Gene Ontology.

al. 2005) databases. To determine whether certain pathways may be preferentially targeted by genetic alterations in human cancer, we compared the genes found to be mutated in Sjöblom et al. (2006) to the constituents of the 825 pathways contained in these four sources. We then determined whether the number of mutations and the distribution of CaMP scores within each pathway is statistically significant using Group CaMP and CaMP GSEA methods. The intersections of the two methods for the four different databases are shown in Figure 4.

In colorectal cancer, 21 pathways were identified to be enriched for mutations from the different pathway databases (Table 3; Supplemental Table 3). Two pathways previously implicated in colorectal tumorigenesis were identified in multiple databases; TGF-beta signaling was identified in three databases and WNT signaling in two. Other signaling pathways contributing to tumorigenesis were also identified, including Insulin signaling, JAK/STAT signaling, MAP kinase signaling, and hedgehog signaling pathways. Two pathways identified were related to the cell cycle and the G_1/S and G_2/M checkpoints. Finally, genes in pathways thought to be important in controlling cell–cell interactions (axon guidance, adherens junctions, and gap junctions) were preferentially mutated in colorectal cancers.

In breast cancer, several known signaling and checkpoint pathways were also identified (Table 3; Supplemental Table 3). These included those involved in AKT signaling, in BRCA1 and BRCA2 repair and cell cycle regulation processes, and in ATM/ATR checkpoint control. Although *TP53* was frequently mutated in each of these pathways, many other genes were also implicated, suggesting that multiple mechanisms may exist for dysregulation of these pathways in breast cancer. Additionally, seven members of the RAN regulation pathway were found to be mutated in breast cancers, while none were mutated in colorectal cancers. The RAN pathway members included proteins involved in nuclear transport such as NUP133, NUP214, NUP98, and KPNA5. NUP98 and NUP214 have been shown to be targets of translocation in several human malignancies (Kau et al. 2004; Nakamura 2005), but no intragenic mutations of these genes have been previously observed in any cancer. In addition, a number of pathways were detected that were related to groups identified using other approaches. These included ABC transporters, which were identified through domain analyses, and the urea cycle pathway, which was identified through the functional group approach.

Integrative analysis

We next integrated the results obtained from the different system-level analyses described above. Although it is not possible to directly compare these disparate groups (e.g., one cannot compare pathways to protein domain groups), one can examine the overlap among the genes belonging to the groups identified through the different analysis modalities. Comparisons of these data, including the set of *CAN*-genes are shown in Figure 5 and in Supplemental Table 4, A and B. Overall, a large fraction (64% and 77%) of the genes in breast and colorectal cancers were uniquely identified by one of the approaches used. These results demonstrate the value of a multidimensional analysis, as each method can identify features that are missed by others. Conversely, the intersecting regions of the different dimensions can point to specific genes that may be of interest because they are implicated by several criteria. For example, in breast cancer, *TP53* and *SPTAN1* were identified to be significant in all five dimensional analyses, and 12 other genes were significant by four methods; in colorectal cancer, five genes lie at the intersection of four of the five analyses, including *TP53*, *EPHA3*, *EPHB6*, *LRP2*, and *IRS4*. Examining the specific groups that identify these and other genes to be enriched may provide insight into new combinations of ways in which these genes may be involved in tumorigenesis.

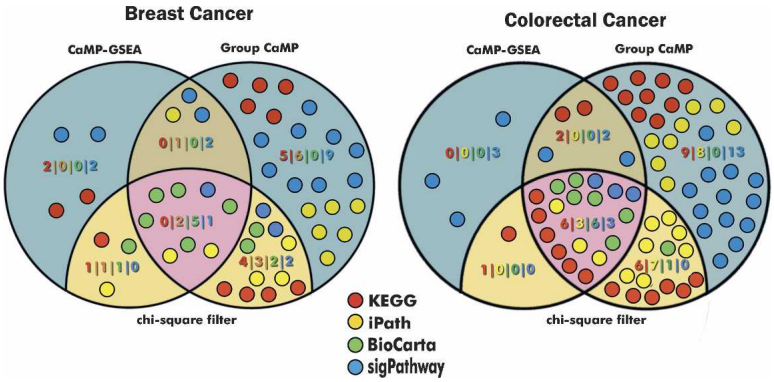

Figure 4. Comparison of mutation enrichment in cellular pathways using complementary statistical approaches. Venn diagrams show the number of pathways identified from four different databases in breast (*left*) and colorectal cancers (*right*) using CaMP GSEA and Group CaMP approaches. Each circle represents one pathway and is colored according to the database it belongs to. Pathways that were enriched for mutations and which were filtered for an increase in the number of genes using the χ^2 test are shown in tan or pink. Blue and dark tan areas represent pathways that were excluded using the χ^2 filter (for additional details, see Methods).

Discussion

Interpreting the large and complex datasets that arise from genome-wide mutational analyses of cancer is challenging. Given the improvements in bioinformatics and sequencing technologies, we expect that many such projects will come to fruition over the next several years. In the first study of this type, Sjöblom et al. (2006) primarily focused on individual genes and attempted to identify those whose mutation frequencies reflected rates that were higher than the passenger mutation frequencies. This initial analysis was unable to discern important genes that were mutated at relatively low frequencies and did not analyze relationships among the mutated genes. The multidimensional analysis performed herein was designed for just these purposes. While the initial study identified individual genes that were mutated, this current analysis identifies pathways, functional groups, and interacting networks that are mutated in colorectal and breast cancers. Although our analyses are at present incomplete and depend on the fidelity and extent of current annotation databases, several conclusions have already begun to emerge from these studies.

The first is that the distribution of mutations observed in the Sjöblom et al. (2006) study is clearly nonrandom. This was initially suggested by the overall number of mutations observed and the prevalence of mutations in specific genes. In this multidimensional analysis, we have identified 51 protein domains, 25 functional groups, and 53 pathways that are enriched for somatic mutations. The Group CaMP and CaMP GSEA approaches used to delineate these gene groups were performed in a rigorous manner, requiring that both the fraction of genes within a particular group and the number or distribution of mutations be higher than what would be expected in the absence of selection. As the Group CaMP studies are sensitive to the passenger mutation prevalence, we performed our analyses using the most conservative estimate available for this prevalence by assuming that all mutations detected in the Sjöblom et al. (2006) study were simply passengers. This approach is clearly an overestimate of the true passenger mutation frequency because it includes mutations of genes that play a causal role in tumorigenesis as well as true passenger mutations. Nevertheless, this process ensured that the groups identified in this study were statistically significant even if the passenger mutation rate was higher than estimated by previous experimental studies.

A second conclusion is that there is substantial value in examining these datasets from different dimensions. Enrichment in protein domains reveals groups of highly related proteins, each of which may be mutated at low levels. Although there is a clear relationship between sequence and function, analysis of enriched functional annotation can allow for abstraction of important biological processes shared by disparate proteins that may not be similar on a sequence level. Examination of protein–protein interactions can provide a more global view of networks that are enriched for mutations. Finally, pathways reveal organizing structures that may not be determined from the other three dimensions. Together, these four complementary views can provide a global view of mutated gene groups and processes.

What are the gene groups and processes that are enriched in these cancer types and what do they tell us about the mechanisms underlying tumorigenesis? For both tumor types, our results pointed to the importance of alterations in intercellular interactions. These included proteins with extracellular domains involved in adhesion (e.g., fibronectin and cadherins), functional groups involved in cell adhesion and extracellular matrix generation and biogenesis, and pathways implicated in cell–cell communication. A multitude of mutated genes are contained in these groups and are delineated in Supplemental Tables 1–3. These observations are generally consistent with the hypothesis that in order for tumor cells to proliferate and invade, they must alter their adhesion dependence to other cells and to the basement membrane and escape control by contact inhibition (Gupta et al. 2007).

The enriched groups and pathways also suggested that certain aspects of intracellular signaling, cell cycle control, and metabolism may be important for tumorigenesis. Two known signaling pathways, involving AKT and ATM/ATR, were enriched in both colorectal and breast tumors, reflecting the important role these play in these tumor types (Vogelstein and Kinzler 2004). However, many of the remaining groups and pathways were specific to one of the tumor types, suggesting that there may be distinct cellular processes underlying breast or colorectal cancer. This is in part exemplified by the dramatic differences of the number and type of interacting mutant proteins present in either breast or colorectal cancers (Fig. 3). For colorectal cancer, such groups included MAD domain-containing proteins, ephrin receptors, metalloproteases, and the TGF-beta and WNT pathways. The latter two of these are consistent with known tumor-related signaling pathways in colorectal cancers (Vogelstein and Kinzler 2004). While the function of metalloproteases and ephrin receptors remains to be further elucidated, one intriguing possibility is that these proteins may be related to the late stage of the tumors examined. All of the colorectal tumors examined in the Sjöblom et al. (2006) study were derived from metastatic lesions, and expression changes (but not mutations) of metalloproteases and ephrin receptors have previously been associated with late stage

Table 3. Cellular pathways identified through mutation enrichment analysis

		Significance[a]			Mutation enrichment					Gene enrichment filter					
Tissue	Database	Group CaMP	CaMP GSEA	Pathway name	No. of mutations	Base pairs sequenced (Mb)	Expected no. of mutations	Group CaMP score	CaMP GSEA score	No. of genes mutated	No. of genes sequenced	Expected no. of genes mutated	Fold change	χ^2 *P*-value	Composite score[b]
Colorectal	Biocarta	●	●	TGF beta signaling pathway	43	416,605	1.5	46.43	>20	5	14	0.7	6.9	6.26×10^{-4}	459.1
		●	●	CTCF: First multivalent nuclear factor	23	226,470	0.8	24.37	9.30	3	10	0.5	5.8	1.22×10^{-2}	195.5
		●	●	Multistep regulation of transcription by Pitx2	33	532,657	1.9	28.83	1.13	3	13	0.7	4.5	2.19×10^{-2}	133.8
		●	●	Cell cycle: G_1/S check point	24	267,949	0.9	24.31	12.30	3	17	0.9	3.4	3.97×10^{-2}	125.0
		●		Inactivation of Gsk3 by AKT	33	425,130	1.5	31.88	0.35	3	21	1.1	2.8	6.30×10^{-2}	89.1
		●	●	NFkB activation by Nontypeable Hemophilus influenzae	9	424,717	1.5	3.10	11.60	3	18	0.9	3.2	4.50×10^{-2}	47.4
	iPath	●	●	WNT signaling pathway Part 1	35	525,457	1.5	31.05	2.08	3	26	1.0	2.9	9.89×10^{-2}	95.9
		●	●	Pleckstrin homology proteins interactions. Part VI	5	251,609	0.7	1.25	9.97	3	9	0.4	8.4	9.63×10^{-3}	93.9
		●		ATM-H2AX-NFBD1 DNA damage response	19	283,949	0.8	16.32	0.34	2	10	0.4	5.0	7.54×10^{-2}	83.6
		●		WNT signaling pathway Part 2	40	1,198,756	3.4	24.68	−0.80	8	74	2.9	2.7	1.38×10^{-2}	64.8
		●		Caspases activation via nuclear import	20	370,506	1.0	15.52	−0.62	3	26	1.0	2.9	9.89×10^{-2}	43.1
		●		Putative integrins pathway Part 2	23	1,114,177	3.1	9.51	−0.52	5	55	2.2	2.3	8.07×10^{-2}	20.5
		●		ATM/ATR regulation of G_2/M checkpoint	21	895,468	2.5	9.56	−1.20	4	41	1.6	2.4	9.39×10^{-2}	20.4
		●	●	A2BR signaling via G-beta/gamma dimer	7	529,597	1.5	1.19	1.05	4	26	1.0	3.9	2.70×10^{-2}	8.6
		●		A2B receptor: action via G-protein alpha s (map 1)	7	543,744	1.5	1.13	0.73	4	32	1.3	3.1	4.85×10^{-2}	5.8
		●		PKA signaling	11	1,256,843	3.5	1.05	−1.59	9	67	2.7	3.4	2.61×10^{-3}	−1.8
	KEGG	●	●	Colorectal cancer	82	1,560,796	4.4	69.21	13.43	12	81	3.2	3.7	2.53×10^{-4}	307.2
		●	●	Pancreatic cancer	47	1,043,695	2.9	36.85	>20	7	61	2.4	2.9	1.57×10^{-2}	163.7
		●	●	Wnt signaling pathway	68	2,448,444	6.9	40.50	4.51	13	136	5.4	2.4	5.37×10^{-3}	107.9
		●	●	Adherens junction	18	1,556,326	4.4	5.03	12.13	7	71	2.8	2.5	3.08×10^{-2}	42.5
		●	●	Long-term depression	26	1,429,187	4.0	11.19	4.08	6	70	2.8	2.2	7.25×10^{-2}	32.8
		●		Axon guidance	31	2,419,478	6.8	9.71	0.48	10	109	4.3	2.3	1.71×10^{-2}	23.4
		●		Gap junction	27	1,874,750	5.2	9.52	0.98	8	93	3.7	2.2	4.15×10^{-2}	22.7
		●	●	TGF-beta signaling pathway	16	1,186,481	3.3	5.27	3.89	7	73	2.9	2.4	3.46×10^{-2}	22.0
		●		Hedgehog signaling pathway	11	901,553	2.5	3.22	0.58	6	51	2.0	3.0	2.21×10^{-2}	11.2
		●		Insulin signaling pathway	25	2,462,009	6.9	5.98	−0.40	9	124	4.9	1.8	7.23×10^{-2}	10.2
		●		Taste transduction	7	653,413	1.8	1.75	−0.33	5	40	1.6	3.1	2.87×10^{-2}	4.4
			●	Cell communication	12	2,583,625	7.2	0.69	1.08	9	97	3.9	2.3	2.16×10^{-2}	4.1
		●		Type II diabetes mellitus	7	845,361	2.4	1.26	−0.32	6	39	1.6	3.9	7.32×10^{-3}	3.6
	SigPathways	●	●	Osteogenesis	19	1,444,220	5.1	4.96	12.90	7	60	3.1	2.3	1.45×10^{-2}	40.3
		●	●	TGF-beta BMP signaling	17	964,252	3.4	6.04	7.27	7	68	3.5	2.0	2.55×10^{-2}	26.5
		●	●	Immunology signaling	57	4,549,048	15.9	16.52	3.23	16	255	13.2	1.2	6.24×10^{-2}	24.0

(continued)

Table 3. *Continued*

		Significance[a]			Mutation enrichment					Gene enrichment filter					
Tissue	Database	Group CaMP	CaMP GSEA	Pathway name	No. of mutations	Base pairs sequenced (Mb)	Expected no. of mutations	Group CaMP score	CaMP GSEA score	No. of genes mutated	No. of genes sequenced	Expected no. of genes mutated	Fold change	χ^2 *P*-value	Composite score[b]
Breast	BioCarta	●	●	Double stranded RNA induced gene expression	20	101,390	0.3	23.87	3.50	3	6	0.2	12.5	8.03×10^{-3}	343.3
		●	●	Chaperones modulate interferon signaling pathway	22	173,661	0.5	22.70	1.74	5	10	0.4	12.5	5.78×10^{-4}	306.6
		●	●	ATM signaling pathway	25	284,806	0.8	22.37	1.50	5	11	0.4	11.4	8.06×10^{-4}	272.3
		●	●	Role of BRCA1, BRCA2 and ATR in cancer susceptibility	24	509,549	1.4	15.68	0.98	4	14	0.6	7.2	1.03×10^{-2}	119.5
		●		Apoptotic signaling in response to DNA damage	20	279,375	0.8	16.13	0.66	3	17	0.7	4.4	7.28×10^{-2}	74.3
		●		Induction of apoptosis through DR3 and DR4/5 death receptors	8	380,314	1.1	1.76	0.15	4	21	0.8	4.8	3.25×10^{-2}	9.1
		●	●	HIV-I Nef: negative effector of Fas and TNF	11	661,120	1.9	2.13	0.14	5	35	1.4	3.6	4.56×10^{-2}	8.1
	iPath	●		Cell cycle regulation by BRCA1	24	424,310	1.5	17.97	0.21	5	28	1.4	3.5	2.18×10^{-2}	62.8
		●	●	Role of BRCA1 and BRCA2 in DNA repair	25	752,288	2.6	12.78	2.92	5	31	1.6	3.1	3.07×10^{-2}	49.0
		●		ATM/ATR regulation of G_1/S checkpoint	24	869,763	3.0	10.59	−0.52	6	41	2.1	2.8	2.73×10^{-2}	28.5
		●		AKT signaling	25	952,072	3.3	10.83	−0.60	7	49	2.5	2.8	1.98×10^{-2}	28.3
		●	●	RAN regulation pathway	12	690,007	2.4	2.05	3.44	7	31	1.6	4.4	2.33×10^{-3}	24.0
			●	NGF signaling pathway	7	441,053	1.5	−0.32	1.83	5	22	1.1	4.4	9.47×10^{-3}	6.7
	KEGG	●		Pancreatic cancer	24	1,035,447	3.6	10.17	−0.75	7	61	3.2	2.2	4.99×10^{-2}	20.9
			●	ABC transporters	9	1,413,582	4.9	0.66	2.38	5	37	1.9	2.6	5.45×10^{-2}	7.9
		●		Urea cycle and metabolism of amino groups	5	331,230	1.2	1.25	−0.39	4	24	1.2	3.2	4.69×10^{-2}	2.8
		●		Regulation of autophagy	5	326,141	1.1	1.27	−0.58	4	28	1.4	2.8	7.07×10^{-2}	1.9
		●		PPAR signaling pathway	10	958,119	3.4	1.55	−0.75	7	69	3.6	2.0	8.08×10^{-2}	1.6
	SigPathways	●	●	Cancer drug resistance	28	1,353,861	3.8	9.98	1.57	8	57	2.3	3.5	1.44×10^{-2}	40.7
		●		Cancer pathwayfind	26	1,268,255	3.6	9.09	0.25	7	55	2.2	3.2	3.26×10^{-2}	29.8
		●		Signal transduction	27	1,390,504	3.9	8.99	−0.28	8	87	3.5	2.3	9.71×10^{-2}	20.1

[a]Pathways wth Group CaMP scores or CaMP GSEA scores >1 were considered significant. The false discovery rates for each of these approaches is estimated to be 10% (for details, see Methods).
[b]The composite score was determined to be the product of the sum of the Group CaMP and CaMP GSEA scores and the fold change of the gene enrichment.

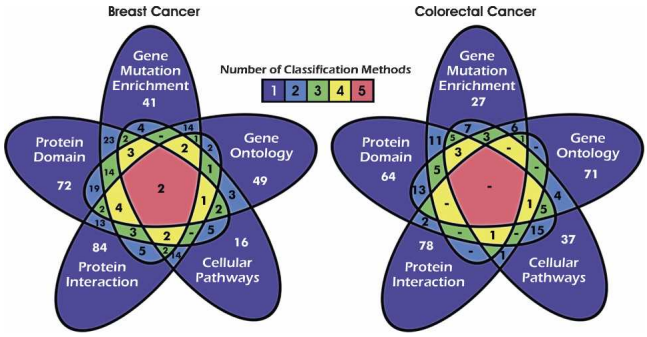

Figure 5. Comparison of genes annotated though different mutation enrichment classification methods. Five-way Venn diagrams (Grünbaum 1975) show the number of genes annotated through the indicated methods for breast and colorectal cancers. The "Gene Mutation Enrichment" set are the *CAN*-genes defined by Sjöblom et al. (2006). Each region indicates the number of genes that are detected by the different analytical methods, and is colored according to the number of methods that identify those genes. The genes detected by each classification method are listed in Supplemental Table 4, A and B.

tumors (Deryugina and Quigley 2006). Breast tumors, on the other hand, demonstrated an enrichment of mutated genes involved in BRCA1 and BRCA2 DNA repair processes, nuclear and cell-surface transporters, urea and arginine biosynthesis, ATPase and GTPase activity, and PPAR signaling pathways. None of these gene groups, except for *BRCA* genes, had been definitively linked to breast tumorigenesis and raise interesting hypotheses about the role of these genes in tumor development and progression. For example, the mutations in the urea pathway affected four proteins (ACY1, ASL, CPS1, and OTC) that impact directly or indirectly on the production ornithine, a key precursor of polyamine synthesis (Casero and Marton 2007). As polyamines have substantial effects on cellular proliferation and apoptosis, mutations of this pathway may represent a novel mechanism of polyamine dysregulation in human cancers.

Finally, the results lead to a deeper understanding of the mutational data and its implications for neoplasia. In the Sjöblom et al. (2006) study, it was noted that there were a great number of genes mutated in each tumor and that the genetic alterations in tumors of the same type were quite heterogeneous. Thus, no two breast cancers shared more than a few mutated genes. This heterogeneity may indeed account for the large biologic differences among breast cancers noted by clinicians, and suggests that a large number of novel therapeutic approaches will be needed to combat these cancers. In contrast, our studies suggest that, though the precise genes mutated in different cancers are heterogeneous, there are a more limited number of genetic groups and pathways through which these genes act. Thus, the complexity at the gene level is likely to be substantially reduced at the pathway level. Future work will be needed to fully elucidate how the gene groups identified herein operate to initiate or accelerate cancerous growth. However, one can at least envision development of therapeutic strategies that inhibit downstream components of a relatively small number of pathways and that would be applicable to a large number of patients. For this vision to be realized, additional sequencing studies as well as more systems biologic analyses, coupled with functional studies of the mutated genes, will all be essential.

Methods

Genomic sequence similarity calculation

The nucleotide and amino acid sequences of all 14,795 CCDS entries were downloaded from NCBI (ftp://ftp.ncbi.nlm.nih.gov/pub/CCDS/current) according to the March 2, 2005 release based on the 35.1 genome annotation build. Using formatdb, we created a CCDS blast database and analyzed each of the CCDS entries using blastp with a minimum *E*-value cut-off of 0.05 and a score of 100. In total, 1639 and 2118 sequences for breast and colon, respectively, were identified and visualized with Cytoscape 2.3.1.

Protein domain comparisons

The InterPro database release 13.0 was downloaded in August 2006 (http://www.ebi.ac.uk/interpro/project_outlines.html) and cross-referenced with the CCDS entries. In total, 12,781 CCDS IDs matched one of 4151 IPR domains. We identified all domains contained within the 1149 mutant genes identified, considering all transcripts of these genes.

Significance of gene sets

Group CaMP

For each group of genes we determined the total number of mutations observed for the tissue of interest, as well as the number of base pairs that were successfully sequenced. We then computed the *P*-value as the probability of a group having at least as many mutations as were observed, given the numbers of base pairs sequenced and the background passenger frequencies, using the binomial distribution in R (http://www.r-project.org). The background passenger frequencies were conservatively estimated as the total numbers of mutations observed in each tumor type in Sjöblom et al. (2006) divided by the total number of base pairs sequenced in the study (i.e., assuming that all of the mutations observed were passengers). The *P*-values were then corrected for multiple comparisons according to the method described by Benjamini and Hochberg (1995) with an FDR of 10%. The Group CaMP score is the negative log of the corrected *P*-value.

CaMP GSEA

We also identified gene sets that were statistically significant in their distribution of CaMP scores (CaMP GSEA) when compared with the whole CCDS set. For each gene, we calculated the CaMP score as described previously (Sjöblom et al. 2006). First, for each of the seven predefined mutational contexts, we computed the probability of observing the number of mutations found using the binomial distribution and the passenger mutation rate. We took the product of these values and divided by the relative rank of this statistic. Using the CaMP score as a metric, we implemented the GSEA algorithm using the R statistical environment. In a list of all the mutated genes sorted by CaMP score, we compared the ranking of the set of genes that contained each domain with those that did not, using the Wilcoxon test, as implemented by the Limma package in Bioconductor (Smyth 2005). Then, the statistical significance of deviation from the null hypothesis of random distribution was calculated and corrected for multiplicity by the Benjamini-Hochberg algorithm (Benjamini and Hochberg 1995) with an FDR of 10%.

For the groups identified by either the Group CaMP or CaMP GSEA approaches, we further focus on those that were also enriched for an increased number of mutant genes. For each

group, we computed the total number of genes observed to be mutated and sequenced, taking into consideration multiple CCDS entries for some genes. For each group, the expected number of mutated genes was calculated to be the product of the number of sequenced genes in the group and the proportion of genes mutated in the entire study. Although this is a post-test filter and not a test in itself, we report a *P*-value calculated using the Pearson χ^2 test in the R statistical environment.

Functional annotation and Gene Ontology

All mutated genes for both colorectal and breast cancers were categorized to general functional categories and visualized using Osprey 1.2.0.

Biological process and molecular function categories were obtained from the GO Consortium website (http://www.geneontology.org). These contained 11,295 biological processes and 7445 molecular functions, as of August 2006. The cross-reference to CCDS entries resulted in 22,705 and 26,430 assignments, respectively. For each GO category, similar calculations were performed for the total number of mutations observed and the total number of base pairs sequenced, as described above for the protein domains. The Group CaMP and GSEA CaMP scores were calculated as described above.

Predicted protein–protein interactions

Predicted protein–protein interactions in humans were downloaded from the Online Predicted Human Interaction Database (OPHID). As of August, 2006, the database contained 49,008 interactions involving 10,682 proteins. CCDS name translation was performed with both RefSeq and SWISS-PROT identifiers as well as via manual curation. Protein–protein interactions between genes mutated in either cancer type were abstracted. In total, 196 and 134 interactions (involving 80 and 59 genes) were identified in breast and colorectal cancers, respectively.

Cellular component data was obtained from Gene Ontology. As of August, 2006, 1802 cellular component terms were available. For network visualization, we first generated the initial network with Cytoscape 2.3.1. Each individual gene was then placed into an appropriate cellular component based on the Gene Ontology data.

Molecular pathways analysis

Pathway assignment data were downloaded from the KEGG (Ogata et al. 1999), Invitrogen iPath (http://escience.invitrogen.com/ipath), BioCarta (http://www.biocarta.com), and sigPathway databases (http://www.chip.org/~ppark/Supplements/PNAS05.html). For KEGG, we used release 39.0 downloaded in August, 2006, which included 41,689 pathways generated from 303 reference pathways. We used NCBI gene IDs to cross-referencing these pathways to CCDS genes, resulting in 7787 assignments. For iPath, we used the online interactive tool to obtain 171 signal transduction and 54 metabolic pathways in August 2006. We identified 4027 assignments of CCDS genes to at least one of the iPath groups. For sigPathway and BioCarta, we used version 1.1.4 from April 2006. We specifically used the 50 pathways identified as humanpaths and 308 pathways identified with BioCarta. In total, 378 and 350 assignments were made to sigPathway and BioCarta. The Group CaMP and GSEA CaMP scores were calculated as described above in the Methods section for the protein domain comparisons to identify cellular pathways that were enriched for mutations in breast or colorectal cancers.

Acknowledgments

This study was supported by The Virginia and D.K. Ludwig Fund for Cancer Research, NIH grants CA 121113, CA 43460, CA 57345, CA105090-03 and CA62924, NSF grant DMS034211, The Pew Charitable Trusts, The Clayton Fund, The Blaustein Foundation, and the NCI Division of Cancer Prevention contract HHSN261200433002C.

References

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. 1990. Basic local alignment search tool. *J. Mol. Biol.* **215:** 403–410.

Apweiler, R., Attwood, T.K., Bairoch, A., Bateman, A., Birney, E., Biswas, M., Bucher, P., Cerutti, L., Corpet, F., Croning, M.D., et al. 2001. The InterPro database, an integrated documentation resource for protein families, domains and functional sites. *Nucleic Acids Res.* **29:** 37–40.

Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., et al. 2000. Gene ontology: Tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.* **25:** 25–29.

Benjamini, Y. and Hochberg, Y. 1995. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. R. Stat. Soc. [Ser A]* **57:** 289–300.

Breitkreutz, B.J., Stark, C., and Tyers, M. 2003. Osprey: A network visualization system. *Genome Biol.* **4:** R22. doi: 10.1186/gb-2003-4-3-r22.

Camon, E., Barrell, D., Lee, V., Dimmer, E., and Apweiler, R. 2004. The Gene Ontology Annotation (GOA) Database–an integrated resource of GO annotations to the UniProt Knowledgebase. *In Silico Biol.* **4:** 5–6.

Casero Jr., R.A. and Marton, L.J. 2007. Targeting polyamine metabolism and function in cancer and other hyperproliferative diseases. *Nat. Rev. Drug Discov.* **6:** 373–390.

Chen, S.T., Choo, K.B., Hou, M.F., Yeh, K.T., Kuo, S.J., and Chang, J.G. 2005. Deregulated expression of the PER1, PER2 and PER3 genes in breast cancers. *Carcinogenesis* **26:** 1241–1246.

Deryugina, E.I. and Quigley, J.P. 2006. Matrix metalloproteinases and tumor metastasis. *Cancer Metastasis Rev.* **25:** 9–34.

Fu, L., Pelicano, H., Liu, J., Huang, P., and Lee, C. 2002. The circadian gene Period2 plays an important role in tumor suppression and DNA damage response in vivo. *Cell* **111:** 41–50.

Giot, L., Bader, J.S., Brouwer, C., Chaudhuri, A., Kuang, B., Li, Y., Hao, Y.L., Ooi, C.E., Godwin, B., Vitols, E., et al. 2003. A protein interaction map of *Drosophila melanogaster*. *Science* **302:** 1727–1736.

Grünbaum, B. 1975. Venn diagrams and Independent Families of Sets. *Mathematics Mag.* **48:** 12–23.

Gupta, G.P., Nguyen, D.X., Chiang, A.C., Bos, P.D., Kim, J.Y., Nadal, C., Gomis, R.R., Manova-Todorova, K., and Massague, J. 2007. Mediators of vascular remodelling co-opted for sequential steps in lung metastasis. *Nature* **446:** 765–770.

Ito, T., Chiba, T., Ozawa, R., Yoshida, M., Hattori, M., and Sakaki, Y. 2001. A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proc. Natl. Acad. Sci.* **98:** 4569–4574.

Jayaraman, L. and Massague, J. 2000. Distinct oligomeric states of SMAD proteins in the transforming growth factor-beta pathway. *J. Biol. Chem.* **275:** 40710–40717.

Jonsson, P.F. and Bates, P.A. 2006. Global topological features of cancer proteins in the human interactome. *Bioinformatics* **22:** 2291–2297.

Kau, T.R., Way, J.C., and Silver, P.A. 2004. Nuclear transport and cancer: From mechanism to intervention. *Nat. Rev. Cancer* **4:** 106–117.

Lee, C.C. 2006. Tumor suppression by the mammalian Period genes. *Cancer Causes Control* **17:** 525–530.

Li, S., Armstrong, C.M., Bertin, N., Ge, H., Milstein, S., Boxem, M., Vidalain, P.O., Han, J.D., Chesneau, A., Hao, T., et al. 2004. A map of the interactome network of the metazoan *C. elegans*. *Science* **303:** 540–543.

Martin, D., Brun, C., Remy, E., Mouren, P., Thieffry, D., and Jacq, B. 2004. GOToolBox: Functional analysis of gene datasets based on Gene Ontology. *Genome Biol.* **5:** doi: 10.1186/gb-2004-5-12-r101.

Nakamura, T. 2005. NUP98 fusion in human leukemia: Dysregulation of the nuclear pore and homeodomain proteins. *Int. J. Hematol.* **82:** 21–27.

Ogata, H., Goto, S., Sato, K., Fujibuchi, W., Bono, H., and Kanehisa, M. 1999. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res.* **27:** 29–34.

Sjöblom, T., Jones, S., Wood, L.D., Parsons, D.W., Lin, J., Barber, T.D.,

Mandelker, D., Leary, R.J., Ptak, J., Silliman, N., et al. 2006. The consensus coding sequences of human breast and colorectal cancers. *Science* **314:** 268–274.
Smyth, G.K. 2005. limma: Linear models for microarray data. In *Bioinformatics and computational biology solutions using R and bioconductor* (eds. R. Gentleman et al.), pp. 397–420. Springer, London, UK.
Subramanian, A., Tamayo, P., Mootha, V.K., Mukherjee, S., Ebert, B.L., Gillette, M.A., Paulovich, A., Pomeroy, S.L., Golub, T.R., Lander, E.S., et al. 2005. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci.* **102:** 15545–15550.
Tian, L., Greenberg, S.A., Kong, S.W., Altschuler, J., Kohane, I.S., and Park, P.J. 2005. Discovering statistically significant pathways in expression profiling studies. *Proc. Natl. Acad. Sci.* **102:** 13544–13549.
Uetz, P., Giot, L., Cagney, G., Mansfield, T.A., Judson, R.S., Knight, J.R., Lockshon, D., Narayan, V., Srinivasan, M., Pochart, P., et al. 2000. A comprehensive analysis of protein-protein interactions in *Saccharomyces cerevisiae*. *Nature* **403:** 623–627.
Vogelstein, B. and Kinzler, K.W. 2004. Cancer genes and the pathways they control. *Nat. Med.* **10:** 789–799.

Received February 23, 2007; accepted in revised form June 28, 2007.