Published in final edited form as:
Genomics. 2008 June ; 91(6): 508–511. doi:10.1016/j.ygeno.2008.03.002.

Functional Classification Analysis of Somatically Mutated Genes in Human Breast and Colorectal Cancers

Thomas W. Chittenden[*,1,2,3], **Eleanor A. Howe**[*,1], **Aedin C. Culhane**[1,2], **Razvan Sultana**[1], **Jennifer M. Taylor**[4], **Chris Holmes**[3,5], and **John Quackenbush**[1,2]

[1] *Department of Biostatistics and Computational Biology and Department of Cancer Biology, Dana-Farber Cancer Institute, Boston, MA, USA*

[2] *Department of Biostatistics, Harvard School of Public Health, Boston, MA, USA*

[3] *Department of Statistics, University of Oxford, Oxford, UK*

[4] *Bioinformatics and Statistical Genetics, Wellcome Trust Centre for Human Genetics, University of Oxford, Oxford, UK*

[5] *MRC Mammalian Genetics Unit, Harwell, UK.*

Abstract

A recent study published by Sjoblom and colleagues [1] performed comprehensive sequencing of 13,023 human genes and identified mutations in genes specific to breast and colorectal tumors, providing insight into organ-specific tumor biology. Here we present a systematic analysis of the functional classifications of Sjoblom's "CAN" genes, a subset of these validated mutant genes that identify novel organ-specific biological themes and molecular pathways associated with disease-specific etiology. This analysis links four somatically mutated genes associated with diverse oncological types to colorectal and breast cancers through established TGF-β1 regulated interactions, revealing mechanistic differences in these cancers and providing potential diagnostic and therapeutic targets.

Keywords

Breast Cancer; Colorectal Cancer; Genomics; Bioinformatics

Introduction

Tumors arise from genetic alterations occurring within a single cell. These are passed to daughter cells that accumulate additional mutations within oncogenes, tumor-suppressor genes, and genomic stability genes [2], which ultimately give rise to tumorigenesis. Although many gene-specific mutations have been discovered in a wide range of cancers, the reference human genome sequence and improved high-throughput sequencing technologies provide the opportunity for an unbiased approach to identification of potentially important mutations. Such

Corresponding Author: John Quackenbush, PhD, Department of Biostatistics and Computational Biology, Dana-Farber Cancer Institute, Sm822, 44 Binney Street, Boston, MA 02115-6084, office: 617 582 8163, fax: 617 582 7760, e-mail: johnq@jimmy.harvard.edu.

[*] These authors contributed equally to this work.

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

an approach was used in a recent study by Sjoblom and colleagues [1] that examined 14,661 transcripts from 13,023 genes (120,839 exons) from the consensus coding sequences (CCDS) database [3] in 11 breast tumors, 11 colorectal tumors, and two normal control samples. A collection of 1,149 potential mutations was detected from which a subset of 236 somatically mutated genes was experimentally validated. Statistical analysis of these validated genes then identified a group of candidate cancer genes (*CAN* genes) that had a higher baseline frequency of mutation than expected by chance: 122 in breast and 69 in colorectal cancers. Grouping organ-specific *CAN* mutant coding sequences into biological themes based on Gene Ontology terms, Sjoblom *et al*. associated these genes with specific cellular processes, including cellular adhesion and motility, signal transduction, transcriptional regulation, transport, cellular metabolism, intracellular trafficking, and RNA metabolism among others. Moreover, they identified significant differences in the mutation spectra of human breast and colorectal cancers, suggesting distinct mutagenic and etiologic pathways.

While enlightening, this qualitative approach to gene functional analysis does not take into account the relative representation of different functional classes associated with the entire collection of genes surveyed. For example, since signal transduction genes are highly represented in the CCDS database, the likelihood of a mutation occurring by chance within these genes is greater than in other, less well-represented classes. With this in mind, we reanalyzed the *CAN* gene dataset from Sjoblom *et al*. to identify biological functional classes and pathways in which greater numbers of genes accumulate mutations than one would expect by chance. Application of a statistically-based categorical representation approach allows one to move beyond looking at individual genes to identify systems, rather than individual genes, that may be associated with disease development and progression. The use of such methods, supplemented by text mining applied to PubMed abstracts associated with those genes, allowed us to identify novel associations of numerous oncological types to colorectal and breast cancers through established TGF-β1 regulated interactions.

Results and Discussion

We analyzed the 69 *CAN* genes found by Sjoblom and associates [1] to be mutated in colorectal cancer, using EASE [4]. EASE uses Fisher’s Exact Test to identify over-represented functional classes relative to the distribution of class assignments for genes in a reference dataset, in this case the CCDS database from which the genes to be sequenced were selected. Functional class assignments included Gene Ontology [5] assignments, chromosome location, phenotype, Pfam domains [6], Swiss-Prot keywords [7], BBID assignments [8], and the GenMAPP [9] and KEGG pathway databases [10]. For *CAN* genes falling within those over-represented classes, we then used Chilibot [11] to perform text mining to delineate associations between the mutant genes and various cancers.

We found 37 *CAN* genes associated with significantly over-represented biological classes. Twenty of these, including *APC*, *TP53* and *TGFBR2*, have been previously associated with colorectal cancer; these largely represent TGF-β signaling, disease mutation, alternative splicing, and proteins containing MH1 and MH2 domains (Table 1). Although none of the MH1-containing proteins were found to have mutations in the active domain, of the MH2-containing proteins, 70% of the *CAN* mutations identified by Sjoblom *et al* were found within the MH2 domain itself. This percentage was not statistically significant according to Chi-square analysis, however, due to the small sample size. The remaining 17, including *CAN* genes associated with Metalloendopeptidase activity and alternative splicing and those containing a Fibronectin type III domain, have not been clinically linked to colorectal cancer as determined by text-mining [11].

The most prominent association revealed in this analysis was the role of TGF-β1 regulation, with 17 of the 37 *CAN* genes having an established relationship to this process (Figure 1A). While mutational inactivation of *TGFBR2* is common in approximately 20–30% of all colorectal cancers [12] and 70% of colorectal cancers with high degree microsatellite instability [13], we find a significant number of additional TGF-β1 regulated genes are also mutated in colorectal cancer, suggesting a much more significant role for this pathway.

Of the 17 TGF-β1 regulated *CAN* genes, *PTPRU* and *RUNX1T1* have not been clinically linked to colorectal cancer. *PTPRU* is implicated in a number of cellular processes including cell growth, cell-cell recognition, cell adhesion, differentiation, mitotic cycle, and oncogenic transformation. The expression of this gene is regulated, in-part, by RAS and upregulated in Jurkat T lymphoma cells [14]. In addition, over expression of *PTPRU* in SW480 cells significantly suppresses cell proliferation and migration, suggesting colorectal carcinomas with mutant *PTPRU* may be more aggressive [15]. Although *RUNX1T1 (ETO)* has not been implicated in TGF-β1 regulation in colorectal cancer, TGF-β1 is a potent endogenous negative regulator of hematopoiesis and the t(8;21)(q22;q22) translocation of this gene, which produces a chimeric protein (AML1-ETO), is one of the most common cytogenetic abnormalities in acute myeloid leukemia [16]. These data implicate aberrant TGF-β1 regulation as a major contributor to disease etiology of colorectal cancer.

This high rate of hits in a single pathway is interesting, particularly since other pathways known to be involved in colorectal cancer were not targeted for mutations in the same manner. For example only six of the sixty-two genes in the WNT/beta-catenin pathway (*SMAD2*, *SMAD3*, *SMAD4*, *TP53*, *TCF7L2*, *APC*) were identified, despite the fact that this pathway is abnormally regulated in 80% of colorectal cancers [17].

Of the 122 *CAN* genes Sjoblom and colleagues [1] identified in breast cancer, we found 24 associated with over-represented biological classes. Only five *CAN* genes involved in cell adhesion molecule activity and GTPase activation have a previously described relationship to breast cancer [11] (Table 1). The remaining 19 *CAN* genes include those linked to JNK activation and proteins containing Spectrin repeat domains. Moreover, all enriched breast cancer terms can be linked to disease-specific cytoskeleton regulation, representing a variety of cellular functions including cell adhesion, migration, proliferation, apoptosis, and differentiation. This suggests cytoskeletal disregulation may be a major contributor to general breast cancer etiology. However, it is well known that breast cancer is a molecularly diverse disease in which subgroups are distinguished by hormone receptor status and gene expression profiles [18]. It is, therefore, unfortunate that more complete data on these tumors are unavailable as it might provide additional insight.

Nevertheless, the available data allows one to draw some interesting comparisons. Unlike colorectal cancer, somatic mutation in breast cancer appears largely TGF-β1 independent. Only 3 of the 24 *CAN* genes identified in our analysis have a known association to TGF-β1 regulation (Figure 1B) and two of these (*COL7A1* and *SPTAN1*) have no clinical link to breast cancer. Type VII collagen (*COL7A1*) defects cause recessive dystrophic epidermolysis bullosa (RDEB), a blistering skin disorder often accompanied by epidermal cancers. Tumor-stroma interactions mediated by collagen VII promote neoplasia in RDEB patients and may contribute to their increased susceptibility to squamous cell carcinoma. *COL7A1* is activated by TGF-β1 via SMAD transcription factors and JUN [19]. Alpha II-Spectrin (*SPTAN1*) is upregulated and associated with tumorigenesis in ovarian cancer. Moreover, TGF-β1 promotes caspase 3-independent cleavage of *SPTAN1*, suggesting mutation of a distinct apoptotic pathway in breast cancer [20].

Our systematic functional classification, comprised of statistical tests for over-represented biological themes and text-mining, provides support for the manually derived themes of recent work by Sjoblom *et al*. [1] and allowed us to identify additional mechanistic insights into the differences between breast and colorectal cancers. In particular, we have identified disease-specific functional classes and somatically mutated molecular pathways that have not been previously reported. We have also found evidence supporting a potentially more significant role for TGF-β1 regulation in colorectal tumorigenesis; a role which highlights mechanistic differences between human breast and colorectal cancers. Furthermore, our analysis identifies four frequently mutated genes (*PTPRU, RUNX1T, COL7A, SPTAN1*) associated with TGF-β regulation that may represent diagnostic and therapeutic targets.

Given the rapid advances in next-generation sequencing technology, we expect to see increasing numbers of sequence-based studies that will expand the catalogue of potentially causative mutations in a wide range of disease states. As we have learned from gene expression studies, functional analysis of the resulting gene lists using now well-established classification systems such as GO can help put the work into an intellectual framework that provides the opportunity for hypothesis generation and mechanistic interpretation. However, such analysis must be applied rigorously to avoid reaching conclusions that reflect trends in the data rather than patterns in the gene set that was sampled.

Materials and Methods

Somatically mutated breast and colorectal candidate cancer genes (*CAN* genes) identified by Sjoblom *et al*. [1] were subjected to a functional category representational analysis using EASE [4] as implemented in MeV [21]. EASE uses Fisher's Exact Test to identify functional classes that appear with a greater likelihood than by chance and calculates associated p-values based on the hypergeometric distribution. Here we analyzed representation for Gene Ontology [5] assignments, chromosome location, phenotype, Pfam domains [6], Swiss-Prot keywords [7], BBID assignments [8], and the GenMAPP [9] and KEGG pathways [10]. In the analysis of GO terms, EASE uses the structure of the GO hierarchy and performs an analysis at each level. One potential limitation of this method is that it identifies as significant those pathways and functional classes that have many genes which have accumulated mutations and does not account for the mutation rates of individual genes.

Chilibot [11] was then used to identify associations between somatically mutated *CAN* genes belonging to the enriched functional classes and disease state. Chilibot is a web-based application that uses natural language processing to search PubMed abstracts for relationships between genes of interest. Each gene is compared with each other gene in the query group and assigned a relationship (stimulatory, inhibitory, neutral, parallel and abstract co-occurrence) based data in the abstract.

Acknowledgments

Thomas Chittenden is supported by an American Heart Association Research Fellowship Award (0525982T). John Quackenbush is supported by grants from the US National Cancer Institute of the National Institutes of Health (R01-CA098522; R01-LM008795) and through funds provided by the Dana-Farber Cancer Institute High Tech Fund.

References

1. 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, Wu L, Liu C, Parmigiani G, Park BH, Bachman KE, Papadopoulos N, Vogelstein B, Kinzler KW, Velculescu VE. The consensus coding sequences of human breast and colorectal cancers. Science 2006;314:268–274. [PubMed: 16959974]

2. Vogelstein B, Kinzler KW. Cancer genes and the pathways they control. Nat Med 2004;10:789–799. [PubMed: 15286780]
3. Database, C. (http://www.ncbi.nlm.nih.gov/CCDS).
4. Hosack DA, Dennis G Jr, Sherman BT, Lane HC, Lempicki RA. Identifying biological themes within lists of genes with EASE. Genome Biol 2003;4:R70. [PubMed: 14519205]
5. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet 2000;25:25–29. [PubMed: 10802651]
6. Pfam. (http://www.sanger.ac.uk/Software/Pfam/).
7. Boeckmann B, Bairoch A, Apweiler R, Blatter MC, Estreicher A, Gasteiger E, Martin MJ, Michoud K, O'Donovan C, Phan I, Pilbout S, Schneider M. The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. Nucleic Acids Res 2003;31:365–370. [PubMed: 12520024]
8. Becker KG, White SL, Muller J, Engel J. BBID: the biological biochemical image database. Bioinformatics 2000;16:745–746. [PubMed: 11099263]
9. Dahlquist KD, Salomonis N, Vranizan K, Lawlor SC, Conklin BR. GenMAPP, a new tool for viewing and analyzing microarray data on biological pathways. Nat Genet 2002;31:19–20. [PubMed: 11984561]
10. Ogata H, Goto S, Sato K, Fujibuchi W, Bono H, Kanehisa M. KEGG: Kyoto Encyclopedia of Genes and Genomes. Nucleic Acids Res 1999;27:29–34. [PubMed: 9847135]
11. Chen H, Sharp BM. Content-rich biological network constructed by mining PubMed abstracts. BMC Bioinformatics 2004;5:147. [PubMed: 15473905]
12. Biswas S, Chytil A, Washington K, Romero-Gallo J, Gorska AE, Wirth PS, Gautam S, Moses HL, Grady WM. Transforming growth factor beta receptor type II inactivation promotes the establishment and progression of colon cancer. Cancer Res 2004;64:4687–4692. [PubMed: 15256431]
13. Ogino S, Kawasaki T, Ogawa A, Kirkner GJ, Loda M, Fuchs CS. TGFBR2 mutation is correlated with CpG island methylator phenotype in microsatellite instability-high colorectal cancer. Hum Pathol 2007;38:614–620. [PubMed: 17270239]
14. Wang B, Kishihara K, Zhang D, Sakamoto T, Nomoto K. Transcriptional regulation of a receptor protein tyrosine phosphatase gene hPTP-J by PKC-mediated signaling pathways in Jurkat and Molt-4 T lymphoma cells. Biochim Biophys Acta 1999;1450:331–340. [PubMed: 10395944]
15. Yan HX, Yang W, Zhang R, Chen L, Tang L, Zhai B, Liu SQ, Cao HF, Man XB, Wu HP, Wu MC, Wang HY. Protein-tyrosine phosphatase PCP-2 inhibits beta-catenin signaling and increases E-cadherin-dependent cell adhesion. J Biol Chem 2006;281:15423–15433. [PubMed: 16574648]
16. Heidenreich O, Krauter J, Riehle H, Hadwiger P, John M, Heil G, Vornlocher HP, Nordheim A. AML1/MTG8 oncogene suppression by small interfering RNAs supports myeloid differentiation of t(8;21)-positive leukemic cells. Blood 2003;101:3157–3163. [PubMed: 12480707]
17. Herbst A, Kolligs FT. Wnt signaling as a therapeutic target for cancer. Methods Mol Biol 2007;361:63–91. [PubMed: 17172707]
18. Perou CM, Jeffrey SS, van de Rijn M, Rees CA, Eisen MB, Ross DT, Pergamenschikov A, Williams CF, Zhu SX, Lee JC, Lashkari D, Shalon D, Brown PO, Botstein D. Distinctive gene expression patterns in human mammary epithelial cells and breast cancers. Proc Natl Acad Sci U S A 1999;96:9212–9217. [PubMed: 10430922]
19. Calonge MJ, Seoane J, Massague J. Opposite Smad and chicken ovalbumin upstream promoter transcription factor inputs in the regulation of the collagen VII gene promoter by transforming growth factor-beta. J Biol Chem 2004;279:23759–23765. [PubMed: 15047696]
20. Brown TL, Patil S, Cianci CD, Morrow JS, Howe PH. Transforming growth factor beta induces caspase 3-independent cleavage of alphaII-spectrin (alpha-fodrin) coincident with apoptosis. J Biol Chem 1999;274:23256–23262. [PubMed: 10438500]
21. Saeed AI, Bhagabati NK, Braisted JC, Liang W, Sharov V, Howe EA, Li J, Thiagarajan M, White JA, Quackenbush J. TM4 microarray software suite. Methods Enzymol 2006;411:134–193. [PubMed: 16939790]

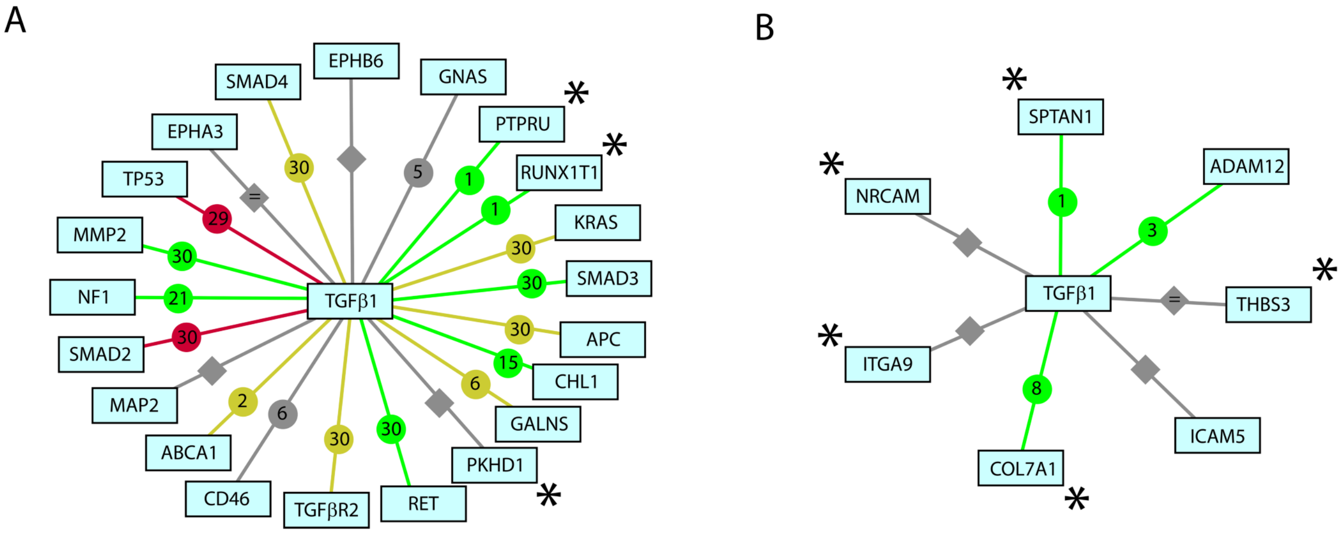

Figure 1.

Chilibot text mining analysis of somatically mutated *CAN* genes in colorectal and breast cancers and their relationship to TGF-β regulation. A. Somatically mutated *CAN* genes in colorectal cancer. B. Somatically mutated *CAN* genes in breast cancer. SYMB Queried terms. ● Interactive relationship (stimulative). ● Interactive relationship (inhibitory). ● Interactive relationship (both stimulative and inhibitory). ● Interactive relationship (neutral). ◆ Non-interactive (i.e. parallel) relationship ◆ Abstract co-occurrence only. Numbers within icons represent PubMed abstracts supporting each association. * Genes have not been clinically implicated in respective cancer states.

Table 1

EASE Analysis

Over-represented functional classes of mutated *CAN* genes found in breast and colorectal cancer. The list of functional annotation classes analyzed include: GO terms for biological process, molecular function, and cellular component; chromosome location; phenotype; protein family (Pfam) domain; Swiss-Prot keywords; Biological Biochemical Image Database (BBID); Gene Map Annotator and Pathway Profiler (GenMAPP) and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways. Here Pop Size is the number of CCDS genes assigned to a particular annotation class. Pop Hits is the number of CCDS sequences assigned to a particular annotation term. List Size indicates the number of mutated genes with assignments in each annotation class and the List Hits is the number associated with each particular term. The Fisher's Exact column lists the *p*-value from Fisher's Exact test and Prob Anal column lists *p*-values corrected for multiple testing via resampling. Genes highlighted in yellow have not been clinically implicated in respective cancer types.

Colorectal Cancer									
Category	**Accession**	**Term**	**List Hits**	**List Size**	**Pop Hits**	**Pop Size**	**Fisher's Exact**	**Prob Anal**	***CAN* Genes**
GO Biological Process	GO:0007167	Enzyme linked receptor protein signaling pathway	8	61	144	10019	2.3×10^{-6}	1.0×10^{-3}	EPHB6, PTPRU, TGFBR2, EPHA3, PTPRD, SMAD4, SMAD3, SMAD2
PFAM domain	PF03166	MH2 domain	3	46	6	5673	9.8×10^{-6}	4.0×10^{-3}	SMAD4, SMAD3, SMAD2
PFAM domain	PF03165	MH1 domain	3	46	6	5673	9.8×10^{-6}	4.0×10^{-3}	SMAD4, SMAD3, SMAD2
GO Molecular Function	GO:0004222	Metalloendopeptidase activity	6	63	86	10197	1.4×10^{-5}	2.0×10^{-3}	ADAM29, ADAMTSL3, ADAMTS18, ADAMTS15, UQCRC2, MMP2
PFAM domain	PF00041	Fibronectin type III domain	6	46	77	5673	3.1×10^{-5}	1.0×10^{-2}	CHL1, EPHB6, PTPRU, CNTN4, EPHA3, PTPRD
GO Biological Process	GO:0007179	TGFbeta receptor signaling pathway	4	61	32	10019	3.9×10^{-5}	1.7×10^{-2}	TGFBR2, SMAD4, SMAD3, SMAD2
GO Biological Process	GO:0007178	Trans-membrane receptor protein serine / threonine kinase signaling pathway	4	61	36	10019	6.3×10^{-5}	2.4×10^{-2}	TGFBR2, SMAD4, SMAD3, SMAD2
SwissProt keyword		Disease mutation	14	44	675	6565	7.9×10^{-5}	5.0×10^{-3}	TP53, TBX22, APC, PKHD1, ABCA1, TGFBR2, ERCC6, GALNS, RET, GNAS, NF1, SMAD4, KRAS, SMAD2
Phenotype		Colorectal cancer	4	17	22	1202	1.7×10^{-4}	1.5×10^{-2}	TP53, APC, TGFBR2, KRAS
GO Biological Process	GO:0007183	SMAD protein heteromerization	2	61	4	10019	2.2×10^{-4}	5.2×10^{-2}	SMAD4, SMAD2
GO Biological Process	GO:0007182	Common-partner SMAD protein phosphorylation	2	61	4	10019	2.2×10^{-4}	5.2×10^{-2}	SMAD4, SMAD2
SwissProt keyword		Alternative splicing	19	44	1310	6565	3.7×10^{-4}	3.8×10^{-2}	TP53, ACSL5, KCNQ5, CD46, APC, ADAM29, PKHD1, RUNX1T1, EVL, GUCY1A2, EPHA3, MAP2, EYA4, PTPRD, GNAS, NF1, KRAS, SMAD2, SFRS6
Phenotype		Pancreatic cancer	2	17	4	1202	1.1×10^{-3}	3.8×10^{-2}	TP53, SMAD4
BBID pathway		84.Ubiquitination_Pathways_Cell_Cycle	2	3	12	291	4.6×10^{-3}	4.1×10^{-2}	APC, CD248
GenMAPP pathway	Human/Gen MAPP.org	Hs_TGF Beta Signaling Pathway	4	15	41	1154	1.4×10^{-3}	2.0×10^{-2}	TGFBR2, SMAD4, SMAD3, SMAD2
Breast Cancer									
Category	Accession	Term	List Hits	List Size	Pop Hits	Pop Size	Fisher's Exact	Prob Anal	*CAN* Genes
GO Molecular Function	GO:0005194	cell adhesion molecule activity	13	109	264	10197	4.4×10^{-6}	2.0×10^{-3}	ADAM12, CDH10, CDH20, COL11A1, COL7A1, ICAM5, ITGA9, NRCAM, RAPH1, TECTA, THBS3, PCDHB15, COL19A1
GO Biological Process	GO:0007257	activation of JNK activity	3	104	5	10019	1.1×10^{-5}	1.0×10^{-3}	DBN1, RASGRF2, MAP3K6

Colorectal Cancer									
Category	**Accession**	**Term**	**List Hits**	**List Size**	**Pop Hits**	**Pop Size**	**Fisher's Exact**	**Prob Anal**	***CAN* Genes**
SwissProt keyword		GTPase activation	5	71	36	6565	3.7×10^{-5}	3.0×10^{-3}	CENTB1, RAP1GAP, CENTG1, RASAL2, STARD8
PFAM domain	PF00435	Spectrin repeat	3	62	11	5673	1.9×10^{-4}	3.8×10^{-2}	SPTAN1, MACF1, SYNE2