Analysis of differential gene expression in colorectal cancer and stroma using fluorescence-activated cell sorting purification

The Harvard community has made this article openly available. Please share how this access benefits you. Your story matters

Citation

Published Version
doi:10.1038/sj.bjc.6604931

Citable link
http://nrs.harvard.edu/urn-3:HUL.InstRepos:28724495

Terms of Use
This article was downloaded from Harvard University’s DASH repository, and is made available under the terms and conditions applicable to Other Posted Material, as set forth at http://nrs.harvard.edu/urn-3:HUL.InstRepos:dash.current.terms-of-use#LAA
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. Culhane1,2, Razvan Sultana1, Jennifer M. Taylor4, Chris Holmes3,5, and John Quackenbush1,2

1Department of Biostatistics and Computational Biology and Department of Cancer Biology, Dana-Farber Cancer Institute, Boston, MA, USA
2Department of Biostatistics, Harvard School of Public Health, Boston, MA, USA
3Department of Statistics, University of Oxford, Oxford, UK
4Bioinformatics and Statistical Genetics, Wellcome Trust Centre for Human Genetics, University of Oxford, Oxford, UK
5MRC 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
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 PT PRU in SW480 cells significantly suppresses cell proliferation and migration, suggesting colorectal carcinomas with mutant PT PRU 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 on 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


6. Pfam. (http://www.sanger.ac.uk/Software/Pfam/).


Figure 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.

<table>
<thead>
<tr>
<th>Category</th>
<th>Accession</th>
<th>Term</th>
<th>List Hits</th>
<th>List Size</th>
<th>Pop Hits</th>
<th>Pop Size</th>
<th>Fisher’s Exact</th>
<th>Prob Anal</th>
<th>CAN Genes</th>
</tr>
</thead>
<tbody>
<tr>
<td>GO Biological Process</td>
<td>GO:0007167</td>
<td>Enzyme linked receptor protein signaling pathway</td>
<td>8</td>
<td>61</td>
<td>144</td>
<td>10019</td>
<td>2.3×10^-6</td>
<td>1.0×10^-7</td>
<td>EPHB6, PTPRU, TGFBR2, EPHA3, PTORP, SMAD4, SMAD3, SMAD2</td>
</tr>
<tr>
<td>GO Biological Process</td>
<td>GO:0007179</td>
<td>TGFbeta receptor signaling pathway</td>
<td>4</td>
<td>61</td>
<td>32</td>
<td>10019</td>
<td>3.9×10^-5</td>
<td>1.7×10^-2</td>
<td>TGFBR2, SMAD4, SMAD3, SMAD2</td>
</tr>
<tr>
<td>GO Biological Process</td>
<td>GO:0007182</td>
<td>Trans-membrane receptor protein serine threonine kinase signaling pathway</td>
<td>4</td>
<td>61</td>
<td>36</td>
<td>10019</td>
<td>6.3×10^-5</td>
<td>2.4×10^-2</td>
<td>TGFBR2, SMAD4, SMAD3, SMAD2</td>
</tr>
<tr>
<td>SwissProt keyword</td>
<td>Disease mutation</td>
<td>14</td>
<td>44</td>
<td>675</td>
<td>6565</td>
<td>7.9×10^-5</td>
<td>5.0×10^-3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Phenotype</td>
<td>Colorectal cancer</td>
<td>4</td>
<td>17</td>
<td>22</td>
<td>1202</td>
<td>1.7×10^-4</td>
<td>1.5×10^-2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>GO Biological Process</td>
<td>GO:0007183</td>
<td>SMAD protein heteromerization</td>
<td>2</td>
<td>61</td>
<td>4</td>
<td>10019</td>
<td>2.2×10^-4</td>
<td>5.2×10^-2</td>
<td></td>
</tr>
<tr>
<td>GO Biological Process</td>
<td>GO:0007184</td>
<td>Common-partner SMAD protein phosphorylation</td>
<td>2</td>
<td>61</td>
<td>4</td>
<td>10019</td>
<td>2.2×10^-4</td>
<td>5.2×10^-2</td>
<td></td>
</tr>
<tr>
<td>SwissProt keyword</td>
<td>Alternative splicing</td>
<td>19</td>
<td>44</td>
<td>1310</td>
<td>6565</td>
<td>3.7×10^-4</td>
<td>3.8×10^-2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Phenotype</td>
<td>Pancreatic cancer</td>
<td>2</td>
<td>17</td>
<td>4</td>
<td>1202</td>
<td>1.1×10^-3</td>
<td>3.8×10^-2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>BBID pathway</td>
<td>84 Ubiquitination Pathways Cell Cycle</td>
<td>2</td>
<td>3</td>
<td>12</td>
<td>291</td>
<td>4.6×10^-3</td>
<td>4.1×10^-2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>GenMAPP pathway</td>
<td>Hs TGF Beta Signaling Pathway</td>
<td>4</td>
<td>15</td>
<td>41</td>
<td>1154</td>
<td>1.4×10^-3</td>
<td>2.0×10^-2</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Breast Cancer

<table>
<thead>
<tr>
<th>Category</th>
<th>Accession</th>
<th>Term</th>
<th>List Hits</th>
<th>List Size</th>
<th>Pop Hits</th>
<th>Pop Size</th>
<th>Fisher’s Exact</th>
<th>Prob Anal</th>
<th>CAN Genes</th>
</tr>
</thead>
<tbody>
<tr>
<td>GO Molecular Function</td>
<td>GO:0005194</td>
<td>cell adhesion molecules activity</td>
<td>13</td>
<td>109</td>
<td>264</td>
<td>10197</td>
<td>4.4×10^-6</td>
<td>2.0×10^-3</td>
<td></td>
</tr>
<tr>
<td>GO Biological Process</td>
<td>GO:0007257</td>
<td>activation of JNK activity</td>
<td>3</td>
<td>104</td>
<td>5</td>
<td>10019</td>
<td>1.1×10^-5</td>
<td>1.0×10^-7</td>
<td></td>
</tr>
<tr>
<td>Category</td>
<td>Accession</td>
<td>Term</td>
<td>List Hits</td>
<td>List Size</td>
<td>Pop Hits</td>
<td>Pop Size</td>
<td>Fisher’s Exact</td>
<td>Prob Anal</td>
<td>CAN Genes</td>
</tr>
<tr>
<td>----------------</td>
<td>-----------</td>
<td>---------------------</td>
<td>-----------</td>
<td>-----------</td>
<td>----------</td>
<td>----------</td>
<td>----------------</td>
<td>-----------</td>
<td>--------------------------------</td>
</tr>
<tr>
<td>SwissProt keyword</td>
<td></td>
<td>GTPase activation</td>
<td>5</td>
<td>71</td>
<td>36</td>
<td>6565</td>
<td>3.7×10^{-5}</td>
<td>3.0×10^{-3}</td>
<td>CENTB1, RAP1GAP, CENTG1, RASAL2, STARD8</td>
</tr>
<tr>
<td>PFAM domain</td>
<td>PF00435</td>
<td>Spectrin repeat</td>
<td>3</td>
<td>62</td>
<td>11</td>
<td>5673</td>
<td>1.9×10^{-4}</td>
<td>3.8×10^{-2}</td>
<td>SPTAN1, MACF1, SYNE2</td>
</tr>
</tbody>
</table>