Meta-analysis of Colorectal Cancer Gene Expression Profiling Studies Identifies Consistently Reported Candidate Biomarkers

Simon K. Chan,1,2 Obi L. Griffith,1,3 Isabella T. Tai,1,4 and Steven J.M. Jones1,2,3

Abstract

Purpose: Elucidation of candidate colorectal cancer biomarkers often begins by comparing the expression profiles of cancerous and normal tissue by performing gene expression profiling. Although many such studies have been done, the resulting lists of differentially expressed genes tend to be inconsistent with each other, suggesting that there are some false positives and false negatives. One solution is to take the intersection of the lists from independent studies. However, often times, the statistical significance of the observed intersection are not assessed.

Methods: Recently, we developed a meta-analysis method that ranked differentially expressed genes in thyroid cancer based on the intersection among studies, total sample sizes, average fold change, and direction of differential expression. We applied an improved version of the method to 25 independent colorectal cancer profiling studies that compared cancer versus normal, adenoma versus normal, and cancer versus adenoma to highlight genes that were consistently reported as differentially expressed at a statistically significant frequency.

Results: We observed that some genes were consistently reported as differentially expressed with a statistically significant frequency (P < 0.05) in cancer versus normal and adenoma versus normal comparisons but not in the cancer versus adenoma comparison.

Conclusion: Our meta-analysis method identified genes that were consistently reported as differentially expressed. A review of some of the candidates revealed genes described previously as having diagnostic and/or prognostic value as well as novel candidate biomarkers. The genes presented here will aid in the identification of highly sensitive and specific biomarkers in colorectal cancer. (Cancer Epidemiol Biomarkers Prev 2008; 17(3):543–52)

Introduction

Colorectal cancer, defined as cancerous growths in the colon, rectum, or appendix, is the third most frequent cancer in both males and females in North America (1). This year, an estimated 20,800 Canadians will be diagnosed with colorectal cancer and ~8,700 will die of it (2). A common area of research interest is the identification of diagnostic biomarkers for early and accurate detection of colorectal cancer (3). Prognostic biomarkers are also being developed to, for example, separate patients who will benefit from adjuvant therapy from those who will not (4) or to determine which patients are at risk for disease recurrence (5). Other studies have focused on understanding cancer progression by identifying differences in gene expression between normal, benign adenoma, and carcinoma stages (6-8).

A common starting point for these studies is the surgical resection of both cancer and normal tissues from patients followed by global expression profiling to determine differentially expressed genes. These studies can result in tens to thousands of such genes, only a small portion of which may actually be of clinical utility. Although an abundance of data comparing the expression profiles of cancerous to normal tissue has been generated, to date, no reliable biomarker has resulted. One explanation for this lack of translational success to the clinic has been the inconsistency in the results of independent studies (1, 9, 10). Explanations for this low overlap include utilization of different tissue resection methods (microdissection, laser capture microdissection, etc.), different expression profiling technologies [cDNA two-channel microarrays, oligonucleotide microarrays, Serial Analysis of Gene Expression (SAGE), etc.], and different analysis methods (multiple correction tests, fold change thresholds, etc.).

Review articles often include lists of genes that have been reported in multiple independent studies. Consistently reported genes are considered to be biologically relevant to colorectal cancer, whereas those reported only sporadically are thought to have resulted from inherent noise or biases in the different platforms and analysis techniques.
methods employed (11). Although these lists are helpful in summarizing the current biomarker candidates, the statistical significance of the level of overlap is usually not considered. One can imagine randomly choosing genes from each expression profiling platform, randomly labeling them as up-regulated or down-regulated, and observing some overlap due to chance alone. Therefore, it would be useful to determine which differentially expressed genes were consistently reported in independent colorectal cancer expression profiling studies with a statistically significant frequency.

To address these challenges, a meta-analysis method was recently developed by our group and applied to published studies of differentially expressed genes in thyroid cancer (12). Such a meta-analysis ignores the differences between studies, such as the expression profiling platform used, and instead focuses on elucidating consistently reported genes. The meta-analysis method involved a vote-counting strategy in which a gene was ranked according to the number of studies reporting its differential expression, the total number of tissue samples used in the studies, and the average fold change. That study resulted in a panel of 12 differentially expressed genes reported at a frequency highly unlikely to have occurred by chance. The panel contained both well-known thyroid cancer markers as well as some uncharacterized genes, showing the ability of the meta-analysis method to highlight novel candidate biomarkers. With these results in mind, the objective of the current study was to apply the meta-analysis method to colorectal cancer to observe whether a statistically significant level of overlap among studies could be observed and to identify promising biomarkers. Also, we improved the meta-analysis method by dividing genes into semiquantitative categories based on the number of tissue samples to highlight genes that may have shown the greatest fold changes but would have been ranked lower by the original method due to fewer tissue samples studied. We curated published lists of differentially expressed genes from 25 independent studies performing global expression profiling to compare colorectal cancer to normal tissue, adenoma to normal tissue, and colorectal cancer to adenoma tissue. Many genes were consistently reported as differentially expressed in multiple studies and this overlap was highly significant. The list of candidate biomarkers we present here will be a valuable resource to the colorectal cancer research community for further studies.

### Materials and Methods

#### Data Collection and Curation
We queried PubMed for colorectal cancer expression profiling studies published between 2000 and 2007. Only studies using tissue samples obtained from surgical resection of cancerous tumors and/or adenomatous polyps were considered. Studies were divided into three comparison types: cancer versus normal, adenoma versus normal, and cancer versus adenoma. We excluded the limited studies that focused on the microsatellite stability of the tissues, specific Dukes stages, or those comparing cancer to cancer samples to determine prognostic biomarkers. In total, differentially expressed genes from 25 independent studies were collected. Twenty-three studies did expression profiling to compare cancer versus normal tissue samples (Table 1), whereas seven and five studies considered adenoma versus normal (Table 2) and cancer versus adenoma (Appendix 6), respectively.

#### Gene Mapping
In the microarray expression profiling studies, differentially expressed genes were represented

<table>
<thead>
<tr>
<th>Reference</th>
<th>Platform</th>
<th>No. genes/features</th>
<th>Up-regulated features (Mapped)</th>
<th>Down-regulated features (Mapped)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Habermann et al. (6)</td>
<td>Hs-UniGEM2 human cDNA microarray</td>
<td>9,128</td>
<td>24 (23)</td>
<td>34 (29)</td>
</tr>
<tr>
<td>Lin et al. (8)</td>
<td>Custom cDNA microarray</td>
<td>23,040</td>
<td>63 (53)</td>
<td>375 (321)</td>
</tr>
<tr>
<td>Buckhaults et al. (19)</td>
<td>SAGE</td>
<td>N/A</td>
<td>153 (106)</td>
<td>246 (201)</td>
</tr>
<tr>
<td>Notterman et al. (17)</td>
<td>Affymetrix Human 6500 GeneChip Set</td>
<td>7,457</td>
<td>19 (19)</td>
<td>47 (45)</td>
</tr>
<tr>
<td>Calamb et al. (45)</td>
<td>Human Atlas Glass 1.0 cDNA microarray</td>
<td>1,090</td>
<td>83 (83)</td>
<td>17 (17)</td>
</tr>
<tr>
<td>Wang et al. (46)</td>
<td>TGS s-4k cDNA microarray</td>
<td>3,800</td>
<td>23 (23)</td>
<td>0</td>
</tr>
<tr>
<td>Croner et al. (20)</td>
<td>Affymetrix HG-U133A</td>
<td>22,283</td>
<td>67 (66)</td>
<td>63 (62)</td>
</tr>
<tr>
<td>Kwon et al. (47)</td>
<td>Macrogen MAGIC cDNA microarray</td>
<td>4,608</td>
<td>77 (77)</td>
<td>45 (44)</td>
</tr>
<tr>
<td>Bertucci et al. (48)</td>
<td>Custom nylon cDNA microarray</td>
<td>8,074</td>
<td>125 (125)</td>
<td>109 (109)</td>
</tr>
<tr>
<td>Ohmachi et al. (49)</td>
<td>Agilent cDNA microarray</td>
<td>12,914</td>
<td>84 (82)</td>
<td>0</td>
</tr>
<tr>
<td>Mori et al. (50)</td>
<td>Human Atlas Glass 1.0 cDNA microarray</td>
<td>1,090</td>
<td>32 (32)</td>
<td>0</td>
</tr>
<tr>
<td>Kim et al. (22)</td>
<td>Oligonucleotide microarray from Compugen/Sigma-Genosys</td>
<td>18,861</td>
<td>272 (271)</td>
<td>216 (216)</td>
</tr>
<tr>
<td>Zou et al. (18)</td>
<td>Custom cDNA microarray</td>
<td>8,000</td>
<td>68 (69)</td>
<td>142 (118)</td>
</tr>
<tr>
<td>Koehler et al. (51)</td>
<td>Atlas Human Cancer 1.2 Array</td>
<td>1,185</td>
<td>31 (29)</td>
<td>14 (13)</td>
</tr>
<tr>
<td>Ichikawa et al. (52)</td>
<td>Custom cDNA microarray</td>
<td>20,784</td>
<td>47 (45)</td>
<td>83 (78)</td>
</tr>
<tr>
<td>Jansova et al. (53)</td>
<td>Human 19K microarrays (Clinical Genomic Centre)</td>
<td>19,201</td>
<td>31 (29)</td>
<td>163 (162)</td>
</tr>
<tr>
<td>Grade et al. (54)</td>
<td>National Cancer Institute oligonucleotide arrays (Operon V2 oligo set)</td>
<td>21,543</td>
<td>1,057 (994)</td>
<td>36 (36)</td>
</tr>
<tr>
<td>Bianchini et al. (55)</td>
<td>Human 19K microarrays (Clinical Genomic Centre)</td>
<td>19,201</td>
<td>76 (76)</td>
<td>12 (12)</td>
</tr>
<tr>
<td>Agrawal et al. (21)</td>
<td>Affymetrix Human 6800 GeneChip Set</td>
<td>7,129</td>
<td>257 (253)</td>
<td>82 (78)</td>
</tr>
<tr>
<td>Sugiyama et al. (56)</td>
<td>Human Cancer Pathway Finder Gene Arrays (Superarray Bioscience)</td>
<td>96</td>
<td>13 (13)</td>
<td>11 (11)</td>
</tr>
<tr>
<td>Kitahara et al. (57)</td>
<td>Custom cDNA microarray</td>
<td>9,216</td>
<td>44 (42)</td>
<td>191 (163)</td>
</tr>
<tr>
<td>Williams et al. (58)</td>
<td>Custom cDNA microarray</td>
<td>9,592</td>
<td>203 (192)</td>
<td>85 (76)</td>
</tr>
<tr>
<td>Takemasa et al. (59)</td>
<td>Custom cDNA microarray</td>
<td>4,608</td>
<td>22 (22)</td>
<td>36 (36)</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td>21,543</td>
<td>1,057 (994)</td>
<td>36 (36)</td>
</tr>
</tbody>
</table>

*Table 1. Twenty-three colorectal cancer versus normal tissue expression profiling studies included in analysis*
by an accession ID, HUGO gene name, or Affymetrix probe ID. The sequence identifier was mapped to the National Center for Biotechnology Information Entrez Gene Identifier (Entrez Gene ID; ref. 13) with the aid of custom-developed Perl scripts and the Clone/Gene ID Converter tool (14). For the SAGE study, updated tag to gene mapping data were obtained from SAGE Genie (15).

**Total Gene Lists.** To estimate the background levels of overlapping studies, we obtained the platform-specific annotation file for each study to identify genes that could potentially be detected as differentially expressed. For commercial platforms, such as Affymetrix and Atlas microarrays, the annotation file was obtained directly from the company Web site. The identifiers in these annotation files were mapped to the corresponding Entrez Gene ID as above to produce a total gene list for each study. Identifiers that could not be mapped to an Entrez Gene ID were ignored. To obtain a total gene list for the SAGE study, all gene names in the tag to gene mapping data from SAGE Genie were mapped to Entrez Gene IDs. For studies that used platforms in which an annotation file could not be obtained, such as the custom cDNA microarrays and some of the oligonucleotide microarrays, an approximation approach was used in which the appropriate number of Entrez Gene IDs was randomly chosen from the combined gene lists from the other platforms.

**Assessment of Significance of Study Overlap Using Simulations.** To determine if the level of overlap among the studies was significant, we did simulations as described previously (12). Briefly, Perl scripts were created to perform Monte Carlo simulations. In each of the 10,000 permutations, the appropriate number of Entrez Gene IDs from the total gene list of each study was randomly chosen and each ID was randomly labeled as “UP” for up-regulated or “DOWN” for down-regulated. We used an “all-or-none” approach in which the level of overlap for a particular gene was only considered if all the independent studies reporting its differential expression agreed on the direction. The level of overlap among studies in each permutation was counted as in the real analysis. On completion of the permutations, a distribution of overlap results from the simulations was determined and a P value was estimated by comparing the overlap from the simulations to the actual level of overlap in the real data. Significance was defined at P < 0.05. Similar to the previous meta-analysis (12), genes were ranked according to three criteria in the following order of importance: (a) level of overlap (that is, listing the same gene as differentially expressed with a consistent direction of change), (b) total number of samples for overlapping studies, and (c) average fold change reported by the studies in agreement. We further subdivided the genes into three categories using a semiquantitative scale: lowest (Q1), moderate (interquartile range), and greatest (values greater than those in Q3) tissue sample sizes. This improvement over the previous published method gives greater importance to the average fold change criteria for ranking genes when total sample numbers are similar.

**Results**

Of the total 8,176 differentially expressed genes reported in the 25 studies (4,273 up-regulated and 3,903 down-regulated), 7,287 (89.1%) could be mapped to an Entrez Gene ID (3,822 up-regulated and 3,465 down-regulated). In the cancer versus normal and adenoma versus normal comparisons, significant overlap was observed. No such significance was seen in the cancer versus adenoma comparison (Table 3), although each individual study identified differentially expressed genes (refs. 6, 8, 17, 19, 35; see Appendix 6).

We present the results from the cancer versus normal comparison as an example. The simulations showed that the amount of overlap in this comparison was highly significant (P < 0.0001), with 573 genes reported as differentially expressed with consistent direction of change in at least two studies (multistudy genes;

### Table 2. Seven colorectal adenoma versus normal tissue expression profiling studies included in analysis

<table>
<thead>
<tr>
<th>Reference</th>
<th>Platform</th>
<th>No. genes/features</th>
<th>Up-regulated genes/features (Mapped)</th>
<th>Down-regulated genes/features (Mapped)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Habermann et al. (6)</td>
<td>Hs-UniGEM2 human cDNA microarray</td>
<td>9,128</td>
<td>20 (19)</td>
<td>38 (35)</td>
</tr>
<tr>
<td>Lin et al. (8)</td>
<td>Custom cDNA microarray</td>
<td>23,040</td>
<td>63 (53)</td>
<td>375 (321)</td>
</tr>
<tr>
<td>Buckhaults et al. (19)</td>
<td>SAGE</td>
<td>N/A</td>
<td>247 (208)</td>
<td>246 (180)</td>
</tr>
<tr>
<td>Notterman et al. (17)</td>
<td>Affymetrix Human 6800 GeneChip Set</td>
<td>7,129</td>
<td>20 (20)</td>
<td>0</td>
</tr>
<tr>
<td>Calamb et al. (45)</td>
<td>Human Atlas Glass 1.0 cDNA microarray</td>
<td>1,090</td>
<td>12 (12)</td>
<td>33 (33)</td>
</tr>
<tr>
<td>Wang et al. (46)</td>
<td>TGS s-4k cDNA microarray</td>
<td>3,800</td>
<td>23 (23)</td>
<td>0</td>
</tr>
<tr>
<td>Lechner et al. (60)</td>
<td>Atlas Human Cancer cDNA microarray</td>
<td>588</td>
<td>15 (11)</td>
<td>9 (5)</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td></td>
<td>400 (346)</td>
<td>701 (640)</td>
<td></td>
</tr>
</tbody>
</table>

### Table 3. Summary of comparisons analyzed for overlap

<table>
<thead>
<tr>
<th>Comparison</th>
<th>Total no. studies</th>
<th>Total no. differentially expressed genes reported (Mapped)</th>
<th>Total no. differentially expressed genes with multistudy confirmation</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cancer vs normal</td>
<td>23</td>
<td>6,537 (5,886)</td>
<td>573</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>Adenoma vs normal</td>
<td>7</td>
<td>1,101 (896)</td>
<td>39</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>Cancer vs adenoma</td>
<td>5</td>
<td>538 (415)</td>
<td>5</td>
<td>0.08</td>
</tr>
</tbody>
</table>

NOTE: The overlap observed in the cancer versus adenoma comparison was not significant (P < 0.05).
observed in the actual data. Simulations never equaled nor exceeded that which was small to visualize. Note that the average overlap values in the intervals of the simulations were not included as they were too small to visualize. Note that the average overlap values in the simulations never equaled nor exceeded that which was observed in the actual data.

Fig. 1). There were 175 multistudy genes that were reported with inconsistent direction of differential expression. Thus, the majority of multistudy genes (76.6%) that were reported as differentially expressed agreed on the direction, even for large numbers of studies.

From the Monte Carlo simulations, an average of 258.30 (95% confidence interval, 258.16-258.45) genes would be expected to have an overlap of 2, whereas the actual data contained 410. An average of 18.37 (95% confidence interval, 18.33-18.42) genes would be expected to have an overlap of 3 compared with 95 in the actual data. For an overlap of 4, the simulation produced 1.14 (95% confidence interval, 1.13-1.15) genes, whereas the actual data contained 30 genes. Overlaps of 5, 6, and 7 were observed in the simulations but with averages of less than one hundredth of a gene. In 10,000 permutations, the simulations never produced an overlap greater than 7, whereas two genes had an overlap of 9 and one gene had an overlap of 11 in the real data. Although the total number of genes with an overlap of 2 was still very significant, we present here only the genes reported by three or more studies, as we deemed these to be the most reliable (Appendices 2-5). Additional information on the results appears in the Appendices (online only).

Discussion

A logical solution to the problem of lack of agreement among expression profiling studies in colorectal cancer is to determine the overlap among many studies using different platforms and observe which genes are consistently reported as differentially expressed. These genes likely show biological relevance to the tumorigenesis of colorectal cancer, as opposed to sporadically reported genes, which may be false positives.

Meta-analyses have been done previously to determine differentially expressed genes in colorectal cancer (1, 9, 10). However, these studies and others usually do not consider whether the level of overlap observed is statistically significant. In the newest version (3.0) of the cancer profiling database Oncomine (16), a meta-analysis tool was implemented to compare results from independent studies. However, Oncomine presently contains raw data for eight colorectal cancer profiling studies, only two of which would qualify for our study (17, 18), because they were the only studies that performed at least one of the three comparisons of interest. As discussed previously, our meta-analysis method is useful when raw data are unavailable for consistent reanalysis, which is usually the case (12). However, one limitation of our method is that a measure of confidence cannot be assigned at the gene level, such as from calculating a true combined fold change or P value. Thus, in order for more powerful meta-analysis methods to be applied to colorectal cancer profiling studies, researchers should be encouraged to make public their raw data so that they may be included in repositories such as Oncomine.

By applying this method to a near comprehensive collection of colorectal cancer expression profiling studies, we were able to determine the genes that were reported with a statistically significant frequency. An extension of the meta-analysis method, we categorized some genes according to their total number of tissue samples as lowest (Q1), moderate (interquartile range), and greatest (values greater than those in Q3) instead of using absolute numbers. This allowed the average fold change criterion to have a greater effect on the gene rank in cases where total sample sizes were similar. In the original version of the meta-analysis method, fold change rarely had any effect on rank.

Overlap Significance Observed in Two of Three Comparisons. We observed that for the cancer versus normal and adenoma versus normal comparisons, genes were consistently reported as differentially expressed and that the level of overlap was statistically significant. The results from the cancer versus adenoma comparison were not significant, suggesting that the number of multistudy genes in the five studies could have been observed due to chance. Determining the significance of overlap among studies provides another filtering step to remove false-positive genes from further consideration. When ignoring the significance of the observed overlap, one may be misled by multistudy genes. For example, without knowledge of the statistical significance, one may reason that the multistudy genes in the cancer versus adenoma comparison are biologically relevant, although this decision cannot be reasonably made because the observed level of overlap may be due to chance alone.

Genes Reported with Inconsistent Direction of Differential Expression. In the cancer versus normal comparison, a total of 748 genes were reported as differentially expressed in at least two independent studies. Although the majority of these genes were reported as differentially expressed in the same direction, 175 (23.4%) genes were not. Of these 175 genes, 132 (75.4%) were reported in two studies, 32 (18.3%) were reported in three studies, 8 (4.6%) were reported in four studies, 2 (1.1%) were reported in five studies, and 1 (0.6%) was reported in six studies. There are many potential explanations for these observed inconsistencies. Firstly, one limitation with such meta-analyses is the
overgeneralization of comparisons. Although every effort was made to ensure that each study included in each of the three comparisons were comparable, there are bound to be inconsistencies due to the lack of relevant clinical data being reported in each of the studies. For example, in the cancer versus normal comparison, SLC26A3 was reported as down-regulated in five studies (8, 17, 19-21) but up-regulated in one study (22). The five studies that reported this gene as down-regulated did not specify the microsatellite status of the colorectal cancer tissue samples being used, whereas the one study that reported the up-regulation of this gene used a mixture of microsatellite stable and unstable tissue samples. Other than microsatellite stability, other clinical features, such as the specific portion of the colon where the tissue samples were taken (9), may affect the direction of differential expression. Thus, due to the lack of these clinical data, it is difficult to determine whether the results of each independent study are truly comparable with each other. Conversely, if these clinical data were more readily available, more specific comparisons, such as microsatellite-stable colorectal tissue samples taken from male patients versus paired normal mucosa, could be done.

A related explanation for why some genes were reported as differentially expressed in an inconsistent direction is the heterogeneity in the tissue samples used. The independent studies experimented on tissue samples taken from vastly different populations, each with different genetic and environmental backgrounds that may contribute to differing expression profiles. Furthermore, the tissue samples used by each study themselves will be heterogeneous compared to each other. To have adequate quantities of tissue to work with, most studies do high-throughput expression profiling on pooled tissue samples, which results in a gene expression signal that is “averaged” across all cells in the samples (21).

However, the expression of a gene in a single cell may be drastically different from this average. Therefore, depending on how the tissue samples were isolated and which ones were pooled together, the genes may be reported as differentially expressed in an inconsistent direction. One of the studies included in the cancer versus normal comparison (21) investigated the feasibility of pooling tissue samples together by plotting the expression signal of all genes in a pooled sample versus the expression signal of genes from one of the samples in the pool. The authors calculated Pearson correlation coefficients and saw that their values ranged from 0.80 to 0.97, suggesting that the pooling of their specific tissue samples maintained patterns of gene expression representative of each distinct tissue sample. Such an analysis should be done in studies using pooled samples to ensure that the pooled versus unpooled results are comparable. Finally, poor study design producing inaccurate results may also explain the presence of these genes. In many cases, these genes were ignored because one lone study reported an inconsistent direction of differential expression, which raises suspicions of the validity of the results of the lone disagreeing study. One concern is that some biologically relevant genes may be omitted due to such a study. Therefore, it may be beneficial to include some genes where the majority of the studies agreed on the direction of differential expression instead of the much more stringent “all-or-none” approach we have used. However, because the majority of these genes (75.4%) were reported in only two studies, including these genes would not alter the identity of the highest-ranking candidates greatly (Tables 4 and 5).

Despite of these inconsistencies, we remind the reader that the majority of the multistudy genes (76.6%) were consistently reported as differentially expressed in the same direction, which is an encouraging result, given that each independent study used diverse experimental techniques and tissue samples.

Literature Review of Cancer versus Normal Candidates. To further assess our results, we performed a literature review of the genes reported by at least seven studies in the cancer versus normal comparison to determine if any have been shown to have diagnostic and/or prognostic utility in colorectal cancer. The most consistently reported differentially expressed gene in our meta-analysis was carbonic anhydrase II (CA2), which was reported as down-regulated in 11 studies. Along with carbonic anhydrase I, CA2 has been shown to have prognostic significance where the expression of both enzymes was related to the metastatic aggressiveness of colorectal cancer (23). Similarly, the potential diagnostic utility of CA2 was shown in a study in which the average level of fecal CA2 in colorectal cancer patients was shown to be significantly greater than those in the control group (24). Immunohistochemistry has been done on colorectal tumor and healthy mucosa tissue to monitor the protein levels of four carbonic anhydrases, among them CA2 (25). That study showed the level of CA2 protein decreased in cancer relative to healthy tissue, thus confirming the transcript based expression profiling results.

Transforming growth factor-β induced 68 kDa (TGFβ1) was reported as up-regulated in nine studies. TGFβ1 is a secreted extracellular matrix protein and was discovered through differential expression analysis of a TGF-β1-treated human lung adenocarcinoma cell line (26, 27). TGFβ1 has also been shown to be strongly induced by TGF-β1 in many other human cell lines (28, 29). Despite the consistent overexpression of this gene, as far as we know, no study has focused specifically on its diagnostic and/or prognostic utility or its role in colorectal cancer tumorigenesis. Overexpression at the protein level has yet to be validated with immunohistochemistry.

IFN-induced transmembrane protein 1 (IFITM1) was also reported as up-regulated in nine studies. IFITM1 has been shown to mediate the antiproliferative properties of the IFN cytokines (30) and was observed to be overexpressed in gastric cancer cells, which resulted in tumor cells being more resistant to natural killer cells and produced a more invasive phenotype (31). As far as we know, immunohistochemistry on human colorectal cancer tissue has not been done for IFITM1 protein; however, reverse transcription-PCR (RT-PCR) was conducted previously on adenomas in a murine model as well as a human colorectal carcinoma cell line, HT29, and elevated expression of IFITM genes (IFITM1, IFITM2, and IFITM3) was observed (32). No further studies have considered the diagnostic and/or prognostic potential of IFITM1 expression in colorectal cancer.

Mal, T-cell differentiation protein-like (MALL), reported as down-regulated in seven studies, is a member of the MAL proteolipid family (33) and encodes an integral
<table>
<thead>
<tr>
<th>Gene name</th>
<th>Description</th>
<th>Studies</th>
<th>Studies with fold change</th>
<th>Total sample sizes</th>
<th>Total sample sizes with fold change</th>
<th>Mean fold change</th>
<th>Range</th>
<th>Validation</th>
</tr>
</thead>
<tbody>
<tr>
<td>TGFβ1</td>
<td>Transforming growth factor-β induced, 68 kDa</td>
<td>9 (8, 19-22, 47, 51, 54, 57)</td>
<td>8</td>
<td>369</td>
<td>329</td>
<td>8.94</td>
<td>1.11-32.00</td>
<td>RT-PCR (8, 19, 51, 54, 57)</td>
</tr>
<tr>
<td>IFITM1</td>
<td>IFN-induced transmembrane protein 1 (9-27)</td>
<td>9 (19, 20, 48, 51-54, 57, 58)</td>
<td>4</td>
<td>351</td>
<td>187</td>
<td>7.52</td>
<td>3.00-12.00</td>
<td>RT-PCR (51, 57)</td>
</tr>
<tr>
<td>MYC</td>
<td>V- myelocytomatosis viral oncogene homologue (avian)</td>
<td>7 (20, 21, 51, 54, 56, 58, 59)</td>
<td>4</td>
<td>329</td>
<td>243</td>
<td>5.02</td>
<td>1.69-7.50</td>
<td>RT-PCR (6, 51, 54, 58) Immunohistochemistry (39)*</td>
</tr>
<tr>
<td>SPARC</td>
<td>Secreted protein, acidic, cysteine-rich (osteonectin)</td>
<td>7 (19-22, 51, 58, 59)</td>
<td>5</td>
<td>244</td>
<td>180</td>
<td>6.30</td>
<td>1.27-15.00</td>
<td></td>
</tr>
<tr>
<td>GDF15</td>
<td>Growth differentiation factor 15</td>
<td>7 (8, 18, 19, 21, 22, 51, 58)</td>
<td>5</td>
<td>230</td>
<td>172</td>
<td>7.42</td>
<td>2.74-10.50</td>
<td>RT-PCR (19, 51)</td>
</tr>
<tr>
<td>CXCL1</td>
<td>Chemokine (C-X-C motif) ligand 1 (melanoma growth-stimulating activity, α)</td>
<td>6 (17, 18, 21, 20, 54, 58)</td>
<td>4</td>
<td>287</td>
<td>229</td>
<td>6.64</td>
<td>2.74-10.50</td>
<td>RT-PCR (18, 21)</td>
</tr>
<tr>
<td>CDC25B</td>
<td>Cell division cycle 25 homologue B (Schizosaccharomyces pombe)</td>
<td>6 (17, 20, 21, 51, 57, 58)</td>
<td>4</td>
<td>256</td>
<td>176</td>
<td>4.93</td>
<td>1.81-9.20</td>
<td>RT-PCR (17)</td>
</tr>
<tr>
<td>HMBG1</td>
<td>High-mobility group box 1</td>
<td>6 (8, 22, 48, 53, 54, 58)</td>
<td>3</td>
<td>264</td>
<td>161</td>
<td>3.27</td>
<td>2.66-3.91</td>
<td>Western blot, immunohistochemistry (61)</td>
</tr>
<tr>
<td>IFITM2</td>
<td>IFN-induced transmembrane protein 2 (1-8D)</td>
<td>6 (17, 18, 21, 20, 54, 58)</td>
<td>4</td>
<td>287</td>
<td>229</td>
<td>6.64</td>
<td>2.74-10.50</td>
<td>RT-PCR (18, 21)</td>
</tr>
<tr>
<td>COL1A2</td>
<td>Collagen, type I, α2</td>
<td>6 (19-22, 53, 59)</td>
<td>3</td>
<td>141</td>
<td>56</td>
<td>7.09</td>
<td>2.74-10.50</td>
<td>RT-PCR (32)</td>
</tr>
<tr>
<td>CCK2</td>
<td>CDC28 protein kinase regulatory subunit 2</td>
<td>5 (17, 21, 22, 51, 54)</td>
<td>5</td>
<td>285</td>
<td>285</td>
<td>4.21</td>
<td>1.79-7.20</td>
<td>RT-PCR (17, 51)</td>
</tr>
<tr>
<td>TOP2A</td>
<td>Topoisomerase (DNA) IIα, 170 kDa</td>
<td>5 (21, 45, 51, 54, 58)</td>
<td>4</td>
<td>277</td>
<td>237</td>
<td>3.61</td>
<td>1.05-5.60</td>
<td>Northern blot, Western blot (62)</td>
</tr>
<tr>
<td>UBE2C</td>
<td>Ubiquitin-conjugating enzyme E2C</td>
<td>5 (20-22, 48, 54)</td>
<td>4</td>
<td>274</td>
<td>229</td>
<td>3.03</td>
<td>1.48-5.00</td>
<td>RT-PCR (63)</td>
</tr>
<tr>
<td>CDH3</td>
<td>Cadherin 3, type 1, P-cadherin (placental)</td>
<td>5 (8, 20, 21, 49, 51)</td>
<td>5</td>
<td>194</td>
<td>194</td>
<td>18.16</td>
<td>2.78-74.00</td>
<td>Western blot (64)</td>
</tr>
<tr>
<td>INHBA</td>
<td>Inhibin, βA (activin A, activin AB α polypeptide)</td>
<td>5 (20-22, 49, 58)</td>
<td>4</td>
<td>198</td>
<td>158</td>
<td>11.05</td>
<td>1.71-37.00</td>
<td>RT-PCR (65)</td>
</tr>
<tr>
<td>SLC12A2</td>
<td>Solute carrier family 12 (sodium/potassium/chloride transporters), member 2</td>
<td>5 (19, 48, 49, 54, 59)</td>
<td>3</td>
<td>208</td>
<td>139</td>
<td>10.58</td>
<td>3.58-15.15</td>
<td>RT-PCR (54)</td>
</tr>
<tr>
<td>MMP11</td>
<td>Matrix metalloproteinase 11 (stromelysin 3)</td>
<td>5 (20-22, 49, 51)</td>
<td>5</td>
<td>208</td>
<td>208</td>
<td>4.22</td>
<td>1.74-5.70</td>
<td>Western blot, immunohistochemistry (66)</td>
</tr>
<tr>
<td>CSE1L</td>
<td>CSE1 chromosome segregation I-like (yeast)</td>
<td>5 (17, 20-22, 48)</td>
<td>4</td>
<td>207</td>
<td>162</td>
<td>3.74</td>
<td>1.14-5.00</td>
<td>None found</td>
</tr>
<tr>
<td>HNRPA1</td>
<td>Heterogeneous nuclear ribonucleoprotein A1</td>
<td>5 (19, 21, 22, 54, 57)</td>
<td>4</td>
<td>243</td>
<td>203</td>
<td>2.89</td>
<td>1.01-4.50</td>
<td>RT-PCR (67)</td>
</tr>
<tr>
<td>CDK10</td>
<td>Cyclin-dependent kinase (CDK2-like) 10</td>
<td>5 (19-21, 45, 49)</td>
<td>5</td>
<td>150</td>
<td>150</td>
<td>13.85</td>
<td>2.66-17.59</td>
<td>None found</td>
</tr>
<tr>
<td>COL3A1</td>
<td>Collagen, type III, α1 (Ehlers-Danlos syndrome type IV, autosomal dominant)</td>
<td>5 (21, 22, 47, 53, 58)</td>
<td>3</td>
<td>178</td>
<td>120</td>
<td>4.31</td>
<td>1.24-9.38</td>
<td>RT-PCR (53)</td>
</tr>
<tr>
<td>COL4A1</td>
<td>Collagen, type IV, α1</td>
<td>5 (20-22)</td>
<td>3</td>
<td>168</td>
<td>126</td>
<td>2.70</td>
<td>1.05-4.00</td>
<td>None found</td>
</tr>
</tbody>
</table>

NOTE: The 22 up-regulated genes reported in at least five independent studies with consistent direction are presented here. Genes reported by five and six studies were further subdivided into semiquantitative categories based on the lowest (Q1), moderate (interquartile range), and greatest (values greater than those in Q3) number of tissue samples to give greater importance to the average fold change criteria for ranking genes when total sample numbers were similar. Validation studies that report a gene as differentially expressed in the opposite direction from that of the meta-analysis are marked with an asterisk.
protein located in glycolipid- and cholesterol-enriched membranes. To the best of our knowledge, its expression at the protein level has not been measured by immunohistochemistry, and diagnostic and/or prognostic utilities have not been studied.

Carcinoembryonic antigen-related cell adhesion molecule 1 (CEACAM1), reported as down-regulated in seven studies, has been shown to be a tumor suppressor in which expression is lost in adenomas and carcinomas. Moreover, the absence of CEACAM1 expression was shown to be correlated with reduced rates of apoptosis in polyps (34). However, a retrospective study did immunohistochemistry on CEACAM1 and showed that 58% of colorectal cancer patients showed an increase in expression (36). It is unclear how the down-regulation of the transcript results in increased CEACAM1 protein expression. Future studies should focus on the half-life of the cancer CEACAM1 transcript to determine if it differs

### Table 5. Down-regulated genes most commonly reported in cancer versus normal expression profiling studies

<table>
<thead>
<tr>
<th>Gene name</th>
<th>Description</th>
<th>Studies</th>
<th>Studies with fold change</th>
<th>Total sample sizes</th>
<th>Total sample sizes with fold change</th>
<th>Mean fold change</th>
<th>Range</th>
<th>Validation</th>
</tr>
</thead>
<tbody>
<tr>
<td>CA2</td>
<td>Carbonic anhydrase II</td>
<td>11 (8, 17, 19-22, 53, 55, 57-59)</td>
<td>7</td>
<td>474</td>
<td>352</td>
<td>-15.51</td>
<td>-56.00 to -2.30</td>
<td>RT-PCR (53, 58)</td>
</tr>
<tr>
<td>MALL</td>
<td>Mal, T-cell differentiation protein-like</td>
<td>7 (17, 19-21, 51, 57, 59)</td>
<td>5</td>
<td>244</td>
<td>180</td>
<td>-5.34</td>
<td>-10.50 to -1.70</td>
<td>None found</td>
</tr>
<tr>
<td>CEACAM1</td>
<td>Carcinoembryonic antigen-related cell adhesion molecule 1 (biliary glycoprotein)</td>
<td>7 (8, 17, 19, 21, 22, 58, 59)</td>
<td>5</td>
<td>222</td>
<td>158</td>
<td>-10.40</td>
<td>-40.00 to -1.38</td>
<td>RT-PCR (17, 58)</td>
</tr>
<tr>
<td><strong>Six studies: greatest sample size</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HSD11B2</td>
<td>Hydroxysteroid (11-β) dehydrogenase 2</td>
<td>6 (8, 17, 20-22, 57)</td>
<td>5</td>
<td>224</td>
<td>184</td>
<td>-4.47</td>
<td>-7.60 to -2.23</td>
<td>Northern blot (68)</td>
</tr>
<tr>
<td><strong>Six studies: moderate sample size</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SLC26A2</td>
<td>Solute carrier family 26 (sulfate transporter), member 2</td>
<td>6 (8, 18, 20-22, 59)</td>
<td>4</td>
<td>190</td>
<td>148</td>
<td>-6.78</td>
<td>-9.09 to -4.04</td>
<td>None found</td>
</tr>
<tr>
<td>FCGBP</td>
<td>Fc fragment of IgG-binding protein</td>
<td>6 (19-22, 48, 57)</td>
<td>4</td>
<td>215</td>
<td>130</td>
<td>-4.88</td>
<td>-7.00 to -1.31</td>
<td>None found</td>
</tr>
<tr>
<td><strong>Six studies: lowest sample size</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ACADS</td>
<td>Acyl-coenzyme A dehydrogenase, C-2 to C-3 short chain</td>
<td>6 (8, 17, 19, 20, 22, 58)</td>
<td>5</td>
<td>168</td>
<td>128</td>
<td>-7.11</td>
<td>-20.00 to -2.00</td>
<td>None found</td>
</tr>
<tr>
<td>CKB</td>
<td>Creatine kinase, brain</td>
<td>6 (19-22, 53, 57)</td>
<td>4</td>
<td>188</td>
<td>130</td>
<td>-3.11</td>
<td>-5.00 to -1.10</td>
<td>Western blot (69)</td>
</tr>
<tr>
<td><strong>Five studies: greatest sample size</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CLI1</td>
<td>Clusterin</td>
<td>5 (17, 21, 47, 53, 58)</td>
<td>3</td>
<td>178</td>
<td>120</td>
<td>-3.83</td>
<td>-5.60 to -1.10</td>
<td>Immunohistochemistry (70) None found</td>
</tr>
<tr>
<td>CES2</td>
<td>Carboxylesterase 2 (intestine, liver)</td>
<td>5 (17, 20-22, 59)</td>
<td>4</td>
<td>186</td>
<td>162</td>
<td>-3.58</td>
<td>-6.30 to -1.15</td>
<td>None found</td>
</tr>
<tr>
<td><strong>Five studies: moderate sample size</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CA1</td>
<td>Carbonic anhydrase I</td>
<td>5 (17, 19, 20, 22, 57)</td>
<td>4</td>
<td>146</td>
<td>106</td>
<td>-36.90</td>
<td>-59.00 to -5.30</td>
<td>RT-PCR (57)</td>
</tr>
<tr>
<td>GPA33</td>
<td>Glycoprotein A33 (transmembrane)</td>
<td>5 (8, 19, 21, 52, 59)</td>
<td>3</td>
<td>131</td>
<td>86</td>
<td>-12.51</td>
<td>-32.50 to -1.70</td>
<td>None found</td>
</tr>
<tr>
<td>KRT20</td>
<td>Keratin 20</td>
<td>5 (17, 19, 21, 22, 57)</td>
<td>4</td>
<td>176</td>
<td>136</td>
<td>-8.31</td>
<td>-20.40 to -1.65</td>
<td>None found</td>
</tr>
<tr>
<td>SELENBP1</td>
<td>Selenium-binding protein 1</td>
<td>5 (19, 20-22, 59)</td>
<td>4</td>
<td>154</td>
<td>130</td>
<td>-2.80</td>
<td>-3.45 to -1.11</td>
<td>Western blot, immunohistochemistry, mass spectrometry (71)</td>
</tr>
<tr>
<td><strong>Five studies: lowest sample size</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CA12</td>
<td>Carbonic anhydrase XII</td>
<td>5 (8, 19, 22, 57, 59)</td>
<td>3</td>
<td>126</td>
<td>62</td>
<td>-4.41</td>
<td>-7.69 to -2.50</td>
<td>Immunohistochemistry (25) None found</td>
</tr>
<tr>
<td>FABP1</td>
<td>Fatty acid binding protein 1, liver</td>
<td>5 (19, 20, 53, 57, 59)</td>
<td>2</td>
<td>116</td>
<td>34</td>
<td>-4.28</td>
<td>-5.56 to -3.00</td>
<td>RT-PCR (57)</td>
</tr>
</tbody>
</table>

**NOTE:** The 16 down-regulated genes reported in at least five independent studies with consistent direction are presented here. Genes reported by five and six studies were further subdivided into semiquantitative categories based on the lowest (Q1), moderate (interquartile range), and greatest (values greater than those in Q3) number of tissue samples to give greater importance to the average fold change criteria for ranking genes when total sample numbers were similar. Validation studies that report a gene as differentially expressed in the opposite direction from that of the meta-analysis are marked with an asterisk.
significantly from the normal CEACAM1 transcript. Furthermore, this study did not observe a relationship between CEACAM1 protein levels and overall survival or disease-free survival in colorectal cancer patients (36).

Secreted protein, acidic and rich in cysteine (SPARC) was reported as up-regulated in seven studies and has been shown by our group to be a putative resistance reversal gene (37). Differentially expressed genes between resistant and sensitive human MIP101 colon cancer cells were determined and SPARC was shown to be consistently down-regulated in the resistant cell lines. Their sensitivity was restored by reexpression of SPARC, suggesting that expression of SPARC has prognostic utility. Immunohistochemistry done on colorectal cancer tissue samples showed increased staining of SPARC protein levels (38). However, another immunohistochemistry study (39) showed down-regulation of SPARC due to methylation of its promoter. Further studies related to the role of SPARC in colorectal tumorigenesis are currently underway in our group.

Growth differentiation factor 15 (GDF15), reported as up-regulated in seven studies, is a member of the TGF-β superfamily. Diagnostic and prognostic utility of GDF15 in colorectal cancer has been suggested by studies that showed increased serum levels of GDF15 in colorectal cancer patients relative to healthy controls (40). These levels increased during disease progression and may have clinical use in the management of colorectal cancer patients (41).

MYC, reported as up-regulated in seven studies in our meta-analysis, is a transcription factor that regulates various processes, such as cell cycle progression, differentiation, apoptosis, and cell motility (42). Immunohistochemistry on MYC has shown that its expression increases during disease progression (43), and when combined with nuclear β-catenin expression, MYC expression was shown to have prognostic utility (44).

In conclusion, the results of this meta-analysis identified genes already shown to have diagnostic and/or prognostic potential in colorectal cancer. Perhaps more interesting are the genes, such as TGFβ1 and IFITM1, which were consistently reported but have yet to be studied specifically as biomarkers. Also, the genes further down the list (that is, those identified as differentially expressed by four, five, six, etc., independent studies) warrant further investigation. Further studies focused on these genes will help in determining a panel of diagnostic and prognostic colorectal cancer biomarkers with sufficient sensitivity and specificity.

Appendix 1: Mapping Success Rate for the Three Comparison Types (Online Only)

The percentage of sequence features that could be mapped to an Entrez Gene ID for each of the three comparison types.

Appendix 2: Up-Regulated Genes Reported in Three or Four Cancer versus Normal Expression Profiling Studies (Online Only)

The 77 up-regulated genes reported by three or four studies were further subdivided into semiquantitative categories based on the lowest (Q1), moderate (interquartile range), and greatest (values greater than those in Q3) number of tissue samples to give greater importance to the average fold change criteria for ranking genes when total sample numbers were similar.

Appendix 3: Down-Regulated Genes Reported in Three or Four Cancer versus Normal Expression Profiling Studies (Online Only)

The 48 down-regulated genes reported by three or four studies were further subdivided into semiquantitative categories based on the lowest (Q1), moderate (interquartile range), and greatest (values greater than those in Q3) number of tissue samples to give greater importance to the average fold change criteria for ranking genes when total sample numbers were similar.

Appendix 4: Up-Regulated Genes Most Commonly Reported in Adenoma versus Normal Expression Profiling Studies (Online Only)

The 23 up-regulated genes reported by two or three studies were further subdivided into semiquantitative categories based on the lowest (Q1), moderate (interquartile range), and greatest (values greater than those in Q3) number of tissue samples to give greater importance to the average fold change criteria for ranking genes when total sample numbers were similar.

Appendix 5: Down-Regulated Genes Most Commonly Reported in Adenoma versus Normal Expression Profiling Studies (Online Only)

The 16 down-regulated genes reported by two or three studies were further subdivided into semiquantitative categories based on the lowest (Q1), moderate (interquartile range), and greatest (values greater than those in Q3) number of tissue samples to give greater importance to the average fold change criteria for ranking genes when total sample numbers were similar.

Appendix 6: Five Cancer versus Adenoma Tissue Expression Profiling Studies Included in Analysis (Online Only)

Appendix 7: Gene Ontology Analysis of Multistudy Genes from the Cancer versus Normal Comparison (Online Only)

Of the 573 multistudy genes, 547 were present in the European Bioinformatics Institute Gene Ontology set of 34,242 annotated genes products. A background set of all genes that were represented at least twice among the platforms used in the cancer versus normal expression profiling studies was used. A total of 24 Gene Ontology terms were found to be statistically overrepresented: 5 biological processes (P), 16 cellular components (C), and 3 molecular functions (F). Number Observed and Total Number show the number of genes from the list found associated with each Gene Ontology term over the total number of genes annotated to that term in Gene Ontology.
References


Meta-analysis of Colorectal Cancer Gene Expression Profiling Studies Identifies Consistently Reported Candidate Biomarkers


Cancer Epidemiol Biomarkers Prev 2008;17:543-552.

Updated version
Access the most recent version of this article at:
http://cebp.aacrjournals.org/content/17/3/543

Supplementary Material
Access the most recent supplemental material at:
http://cebp.aacrjournals.org/content/suppl/2008/03/21/17.3.543.DC1

Cited articles
This article cites 70 articles, 23 of which you can access for free at:
http://cebp.aacrjournals.org/content/17/3/543.full#ref-list-1

Citing articles
This article has been cited by 4 HighWire-hosted articles. Access the articles at:
http://cebp.aacrjournals.org/content/17/3/543.full#related-urls

E-mail alerts
Sign up to receive free email-alerts related to this article or journal.

Reprints and Subscriptions
To order reprints of this article or to subscribe to the journal, contact the AACR Publications Department at pubs@aacr.org.

Permissions
To request permission to re-use all or part of this article, contact the AACR Publications Department at permissions@aacr.org.