Published in final edited form as:
Proteomics Clin Appl. 2009 April ; 3(4): 473–485. doi:10.1002/prca.200800074.

Integrative Analysis of Cancer Pathway Progression and Coherence

Ertugrul Dalkic[1,2,3], **Daniel Elwin Walter Nash**[1,2,7], **Mohammad Kasim Fassia**[1,2,8], and **Christina Chan**[*,1,2,3,4,5,6]

[1] Center for Systems Biology, Michigan State University, East Lansing, MI 48824, USA

[2] Cellular and Molecular Biology Lab, Department of Chemical Engineering and Materials Science, Michigan State University, East Lansing, MI 48824, USA

[3] Cell and Molecular Biology Program, Michigan State University, East Lansing, MI 48824, USA

[4] Department of Biochemistry and Molecular Biology, Michigan State University, East Lansing, MI 48824, USA

[5] Department of Computer Science and Engineering, Michigan State University, East Lansing, MI 48824, USA

[6] Department of Chemical Engineering and Materials Science, Michigan State University, East Lansing, MI 48824, USA

[7] Department of Mathematics, Michigan State University, East Lansing, MI 48824, USA

[8] Lyman Briggs College of Science, Michigan State University, East Lansing, MI 48824, USA

Abstract

We analyzed the cancer pathways in the KEGG (Kyoto Encyclopedia of Genes and Genomes) database. The database provides a collective of signaling pathway members involved in cancer progression. However, the KEGG cancer pathways, unlike signaling pathways, have not been analyzed extensively with gene expression and mutation data. We transformed the colorectal cancer pathway into discrete X and Y scales and analyzed the relative expression levels of adenoma and carcinoma samples as well as the distribution of mutation targets. The X scale corresponds to the downstream location in a pathway, whereas the Y scale corresponds to the stage of the tumor. The gene expression values of the early stage pathway members are expressed significantly higher than the rest of the pathway members in colorectal adenoma tissues. The colorectal cancer pathway shows some degree of coherence in the carcinoma samples. The correlated gene pairs responsible for the coherence of the colorectal cancer pathway in the carcinoma samples are supported, in part, by the literature and may suggest novel regulatory associations. Finally, there are more mutation targets in the nucleus as well as the late tumor stages of the KEGG colorectal cancer pathway.

Keywords

biological pathways; cancer; systems biology

[*]Address correspondence to: Christina Chan, Michigan State University, Department of Chemical Engineering and Materials Science, 2527 Engineering Building, East Lansing, MI 48824, Tel.: [517] 432-4530, Fax: [517] 432-1105, krischan@egr.msu.edu.

The authors declared no conflict of interest.

1 Introduction

Cancer is a complex disease, with many subtypes, affecting various tissues, and according to the severity of the abnormality, giving rise to different stages and classifications, such as carcinoma, sarcoma, primary, etc. Physiological and genetic studies identified different stages in the progression of the various types of cancers. For instance, colorectal tumorigenesis begins with normal epithelium which proceeds through stages of hyperproliferative epithelium, early adenoma, intermediate adenoma, late adenoma, carcinoma, and metastasis [1]. Genetic alterations, such as mutations or deletions of genes, accumulate as the cancer progresses to the next, more severe and proliferative stage. The accumulation of changes (e.g., mutations, deletions, etc.) is a key factor in tumor progression. For example, a more severe stage will have a higher probability than a less severe stage of having more mutations [1]. On the other hand, the specific number and the identity of the genetic alterations, such as point mutations, amplifications, or deletions in a stage, are not determining factors, but rather are general features that characterize the severity of a stage, and varies from one cancer to another [1].

Omics technologies, such as cDNA and oligonucleotide arrays, and comparative genome hybridization, have dominated tumor characterization [2]. Genome level analysis have been applied successfully to differentiate between the different stages of cancer, and revealed differentially expressed genes and genomic alterations that play important roles in the development of cancer [2]. More recently, gene set and pathway centric analysis of cancer progression have gained popularity. These analyses confirmed results currently known about the role of cell cycle in cancer development, as well as produced novel findings, i.e. the involvement of the ERBB4 gene in primary prostate cancer [3]. Approaches that integrate gene expression and pathway analysis of colorectal cancer have proved useful in finding potential prognostic and diagnostic markers, as well as therapeutic targets [2].

Biological pathways consist of molecular interactions or other biochemical events among a group of proteins, genes, and other chemicals. They are used to characterize metabolism, signal transduction, specific cellular processes, diseases, or drug mechanisms. The Kyoto Encyclopedia of Genes and Genomes (KEGG) database contains pathway information on the progression of different types of cancer, as combinations of signaling pathways (http://www.genome.ad.jp/kegg/pathway.html#disease) [4,5,6]. Signaling pathways capture the molecular interactions and reactions that take signals from the outside to the nucleus of the cell, where transcriptional regulation occurs. Signaling pathways, e.g. the MAPK, Wnt, TGF-beta signaling, are well studied in the context of cell proliferation [7]. Cancer is the only human disease for which pathway information of the disease progression is available, i.e., at different stages of the cancer, from normal tissue to the advanced tumor phase in KEGG [4,5,6]. KEGG cancer pathways are different in that they contain members of different signaling pathway members within a single pathway. This is because they contain information on various stages of the cancer and the various stages involve different signaling pathways. The pathway information contained in KEGG, i.e., the genes in each stage of the disease progression that are genetically altered in the colorectal cancer pathway, is derived from the literature [8,9]. In the KEGG colorectal cancer pathway, the Wnt pathway members are at the upper positions of the image, corresponding to normal or initial stages of the cancer progression, which is supported by the literature [8,9,10]. Constant activation of Wnt signaling targets occurs in adenomatosis polyposis coli (APC) or beta-catenin mutations, as well as upon phosphorylation by glycogen synthase kinase-3-beta. These events lead to early dysplastic lesions, which is an early event in colorectal cancer [10]. Similar support exists in the literature to suggest that genes, i.e., DCC and KRAS, should be below the Wnt pathway members since their alterations occur later in the disease progression. Furthermore, DCC is located below KRAS in the KEGG pathway because

DCC alterations are believed to occur later than KRAS alterations in the development of colorectal cancer [11]. Finally, TGF-beta pathway members are located in the lower portions of the pathway image because the literature suggests that their alterations occur during more advanced stages of the cancer [12]. The significance of the KEGG cancer pathways is the integration of the cancer stages with signaling pathways. Although the integration of signaling pathways with genome level expression data has been widely performed, it has yet to be realized with cancer pathways.

Using pathway information to understand genome level expression data has been extensively applied [13,14,15]. The approach integrates *a priori* knowledge of a gene's functional role with expression data to detect for concerted expression changes in a set of genes responsible for producing a phenotype [13]. Pathway-centric analysis of tumor microarray data has been successfully applied to identify signaling pathway members [16]. IL-1 and ER-induced pathways were found to be significantly coexpressed in breast cancer data. In addition to analyzing the entire dataset, analysis of individual samples or a subset of the data identified significant pathway activities that were relevant to the biological context of the tissue or organ. For example, pathway analysis uncovered the expected association of estrogen-induced pathways within a group of clinical breast cancer data [16], and signaling and metabolic pathways involved in the development of type 2 diabetes [17]. On the other hand, gene expression level analysis *within* pathways is an area that mostly has been ignored. Some studies have analyzed the relationship among the members of protein complexes or pathways in terms of their gene expression levels [18,19,20,21]. Protein complexes, such as ribosome and proteasome, show significant correlation in their gene expression levels [18]. In addition, the cis-element profiles are highly similar for members within a signaling pathway, such as the KEGG apoptosis pathway, and functionally related interacting proteins (i.e., protein complexes). This suggests that a strong relationship in the gene expression levels between members of these pathways should exist [19]. Coherence is a measure of the level of correlation among a group of genes. A coherent group of genes may share similar regulation of their gene expression levels. Indeed, genes in the same pathway with similar functions have been shown to be coherent as compared to a random group of genes from the genome [20].

Previous studies that integrate gene expression data with pathway information have not incorporated the dimensionality of the pathways. Most studies have focused on the members of the pathways. Thus far, there has been one study of pathways that incorporated the position in the pathways in the analysis of the genome level data. They developed a statistical impact analysis that used the pathway position to calculate the significance of the pathway [21]. With this approach they identified the Focal Adhesion Pathway as a significant pathway for lung cancer, which was not found using classical approaches, such as gene set enrichment and gene ontology analyses, thereby enhancing the information content extracted from analyzing genome level data. Impact analysis considers expression alteration in receptors, such as integrin, receptor tyrosine kinase (RTK), and the receptor ligand vascular endothelial growth factor (VEGF), as important parameters in the analysis, since they affect the downstream molecules in the pathway. The Focal Adhesion Pathway, on the other hand, was not found to be significant using classical approaches because the other genes in the pathway were not significantly altered. Classical approaches analyze pathways as a whole without special emphasis on receptor molecules or other position(s) in a pathway.

In this study, we capitalized upon the progressive nature of the cancer disease captured in the biological knowledge represented in the KEGG cancer pathways. We analyzed the gene expression levels, coherence, and mutation target data of the pathway members to determine if there is a significant relationship or correlation within any group of pathway members

from the rest of the pathways. Analyzing the KEGG colorectal cancer pathway with microarray data identified that different parts of the pathways were up-regulated or coherent at the mRNA level, at the different stages of cancer progression, i.e., adenoma vs. carcinoma. Since the KEGG cancer pathways integrate different signaling pathways of the various cancer stages, unlike classical signaling pathways, we analyzed the coherence of the colorectal cancer pathway, and found the carcinoma expression data was more coherent than the normal or adenoma data. In addition, mutation targets were found to be localized primarily in the nucleus of the cell and concentrated at the later stages of the cancer.

2 Materials and methods

2.1 Pathway data

We used KEGG as our source of pathway information [4,5,6]. We focused on the cancer, apoptosis, oxidative phosphorylation and proteasome pathways. We collected a list of human genes for these pathways. For each protein/gene in the pathways we used all the different homologues provided by KEGG (Supplementary Tables 1–17).

2.2 X/Y scale

The X scale represented the direction from receptor to nucleus. The Y scale (analyzed only for the colorectal cancer pathway) represented the direction from normal tissue to advanced tumor or metastasis (Figure 1). We ignored the sub-pathway distinction (such as Chromosome Unstable Pathway and Microsatellite Unstable Pathway in colorectal cancer pathway) (Figure 1). We generated 3 X-dependent and 4 Y-dependent groups in the KEGG colorectal cancer pathway, and 2 X-dependent groups in the KEGG apoptosis pathway. For X, receptor ligands and receptors are designated by a value of 1. If there is a nuclear distinction in the pathway; molecules in the cytosol are designated by a value of 2, and molecules in the nucleus are designated by a value of 3, otherwise molecules in the cytoplasm are designated by a value of 2. For Y (analyzed only for the KEGG colorectal cancer pathway), the first stage of the pathway, which signified the first initial molecular events (in the colorectal cancer pathway, the first stage is represented by the transition from normal epithelium to early adenoma) is designated by a value of 1 and the next stages in the cancer progression are designated by values of 2, 3, 4 (the last stage). The last stage is defined by the latest events in the progression of the disease in the KEGG pathway. For example, in the colorectal cancer pathway, the last stage is represented by the transition from late adenoma to carcinoma. The biomolecules involved in the transition to the different stages provided by KEGG are supported by the literature, as discussed in the Introduction above. The KEGG information was used as is, unless the stage is an insignificant one. For example, dysplastic aberrant crypt foci stage was ignored in our analysis of the colorectal cancer pathway, because this stage covered a very small part of the pathway and it was unclear which molecules were associated with this stage. In the colorectal cancer pathway, normal epithelium to early adenoma is designated by a Y value of 1 and includes proteins such as, Frizzled, glycogen synthase kinase-3-beta (GSK-3β), adenomatosis polyposis coli (APC), T cell factor/lymphoid enhancer factor (TCFLEF), Survivin, etc. Early adenoma to intermediate adenoma is designated by a Y value of 2 and includes proteins such as, RTK, K-Ras, protein kinase B (PKB), extracellular signal regulated protein kinase (ERK), C-Fos, etc. Intermediate adenoma to late adenoma is designated by a Y value of 3 and includes proteins such as, deleted in colon cancer (DCC), caspase 3 (CASP3), human mutL homolog 1 (hMLH1), etc. Late adenoma to carcinoma is designated by a Y value of 4 and includes proteins such as cytochrome c (Cytc), p53, etc. Since our analysis focused on the potential of the genes in contributing to the progression of cancer, we used the larger value of Y for a gene if there were more than one value associated with the gene. For consistency, the same approach was taken for the X value, i.e., the larger of the two values was used. For example,

if a protein is present in more than one location, the larger value is assigned, for instance, CyclinD1 is present in two Y locations, 1 and 2, but was assigned a Y value of 2, and transforming growth factor beta receptor 2 (TGFBR2) is located in two X locations, 1 and 3, but was assigned a X value of 3 (Figure 1).

2.3 Expression data

Normalized microarray data (GSE4183 and GSE8671) were downloaded from the NCBI GEO database (http://www.ncbi.nlm.nih.gov/geo/) GSE4183 includes normal colon tissue, colon adenoma and colon carcinoma gene expression datasets for several biopsy samples obtained from individuals. The GSE8671 dataset includes the whole genome mRNA expression level for 32 colorectal adenomas paired with the normal mucosa from the same individuals. In these large-scale datasets, we focused on the expression values of the pathway members, as oppose to analyzing all the genes. Both datasets contained expression values for all the gene members in the colorectal cancer pathway. If there were more than one value for a gene in a sample (same column), the mean values were used. We combined the normal and the adenoma samples from both datasets, to obtain a set of expression values for normal, adenoma and carcinoma samples. We calculated expression ratio for adenoma (or carcinoma) by dividing the average value of adenoma (or carcinoma) samples to the average value of normal samples. For adenoma, the ratio was calculated by dividing the average of the combined adenoma values to the average of the combined normal values. Since carcinoma data is present only in GSE4183, the average of the carcinoma values in GSE4183 were divided by the average of the normal values in GSE4183. For the calculation of carcinoma/adenoma ratio, the carcinoma ratio was divided by the adenoma ratio for each gene (Supplementary table 1).

2.4 Drug and mutation targets

We collected drug information from the National Cancer Institute web page (http://www.cancer.gov/cancertopics/druginfo/alphalist). The drugs, cancers for which they are used, and their targets are listed in Supplementary Table 19. We limited the drug list to ones which targeted a cellular signaling protein. Drugs that targeted molecules with a general role in the cell were excluded; for example, Bortezomib, a drug approved for leukemia, was excluded because it targeted the proteasome. We collected mutation target information from two sources. The first source was the Cancer Gene Census of the Cancer Genome project [22] (http://www.sanger.ac.uk/genetics/CGP/Census/). We used only the cancer types provided by this dataset. For some mutation targets, such as p53, “others” was mentioned in addition to the cancer type listed, therefore we were not able to include the cancer types referred to as “others” in the analysis. The second source was a list of genetically altered genes provided by KEGG for each pathway. We combined these two sources of data to obtain a list of mutation targets for each cancer pathway. The list from the two sources did not overlap, for example, BRAF was identified as a mutation target only in the Cancer Gene Census dataset, while TGFBR2 was identified as a mutation target only in the KEGG dataset. The list of members of the colorectal cancer pathway, the tumor expression ratio values, X/Y values, and mutation and drug target information are provided in Supplementary Table 1. If the gene is a mutation or drug target, it was denoted by 1, if not, it was denoted by 0. Only mutation and drug target information for the list of members of the remaining cancer pathways are provided in Supplementary Tables 1, 5–17. For mutation targets we calculated the frequency as the number of mutation targets in a group divided by the total number of genes in the group.

2.5 Statistical Analysis

Analysis of groups of genes or members of a pathway was performed for their coherence in the normal, adenoma, and carcinoma tissues. We calculated the Pearson’s correlation

coefficient for the expression values of every pair of genes in both datasets. We selected the same number of random genes from the entire microarray dataset, whether the genes belonged in the pathway or not, and performed the same calculation for this random group. For the range between 0–1, at increments of 0.01, we calculated the fraction of pairs with a correlation coefficient for both the real groups (pathways, subgroups inside pathways) and the random groups. For each gene, we performed the randomizations 1000 times for the pathways and 100 times for the subgroups within the pathways, and calculated the fraction of random pairs with a correlation coefficient threshold of 0.5. In addition to the colorectal cancer pathway, we analyzed the apoptosis pathway because it was previously shown to be coherent in colorectal cancer [20]. We also analyzed the oxidative phosphorylation and proteasome pathways which were shown to be coherent both in normal and cancer tissues [20]. The oxidative phosphorylation pathway has more genes than the colorectal cancer and apoptosis pathways, while the proteasome has fewer genes. Therefore any size effect of the pathway should be accounted for. For each X/Y dependent subgroup, randomizations of the same size were performed and whether the subgroups differed significantly from random was determined at a correlation coefficient threshold of 0.5. For the analysis of expression ratios, we performed t-test on groups of 2 (i.e., comparing gene expression profiles of pathway members in group Y=1 to members in the rest of the pathway, namely Y=2, 3, 4). For more than two groups (i.e., comparing gene expression profiles of pathway members in groups X=1, X=2, and X=3), we used ANOVA.

2.6 Selection of the Colorectal Cancer Pathway

We combined information of the mutation and drug targets with information on the pathway to analyze the coverage, namely, the number and distribution of mutation and drug targets in the pathways for several types of cancer (Supplementary Table 1, 5–17). Although, Acute Myeloid Leukemia (AML) had the highest known number of mutation targets, 82, the KEGG pathway did not include many of their targets as compared to thyroid and colorectal cancers, which had 25 and 20 mutation targets, respectively. 11 mutation targets are present in the KEGG pathway for AML, while thyroid and colorectal cancers had 14 mutation targets in their KEGG pathway. Therefore, the KEGG pathway provided a higher coverage of the mutation targets for thyroid and colorectal cancers. Glioma also had a high representation of their mutation targets, namely 11 of 11 total targets are present in the KEGG pathway. In contrast to mutation targets, there were a limited number of drug targets shown in the KEGG pathway for Glioma; therefore AML and Glioma were not analyzed extensively. Similarly, thyroid cancer had a very low number of total pathway members and thus was also not analyzed.

We considered the high number of pathway members and mutation targets, availability of large-scale expression data, and the clarity and complexity of the stage information of the KEGG pathway in selecting the cancer type(s) for this integrative analysis. For example, the presence of the many stages in AML without a clear distinction of the pathways between the different stages of this cancer made it inappropriate for this analysis. Similarly, the simplicity of the glioma pathway (presence of only one stage, or presence of too many proteins in a stage and very few proteins in the other stages) and thyroid cancer pathway (too few pathway members) made them inappropriate for this integrative pathway analysis. Therefore given the availability of the different types of data (i.e., mutation and drug targets, and expression data) that is required and the clarity and the size of the KEGG pathway, we focused on the colorectal cancer pathway to demonstrate the feasibility of this integrative approach.

3 Results and discussion

3.1 Gene expression ratio analysis of dimensional grouping of the colorectal cancer pathway

We analyzed the tumor expression levels of the members of the KEGG colorectal cancer pathway. We investigated whether the pathway members are differentially expressed with respect to their X (*cellular location*), and Y (*stage of the tumor*) values in the KEGG pathway, see Materials and Methods for details. The colorectal cancer pathway in KEGG provides some of the molecular events that underlie the progression of cancer, from normal to carcinoma. We assessed whether a relationship existed between the progression of colorectal cancer and the gene expression levels of the members of the colorectal cancer pathway. The normal and the adenoma samples from both datasets (GSE4183, GSE8671) were combined and the carcinoma dataset came from GSE4183.

Analyzing the genes that were highly differentiated among the pathway members in the carcinoma and adenoma datasets, we found AXIN2 and FZD3 genes were the most down-regulated in the carcinoma and the most upregulated in the adenoma samples. These genes had the lowest carcinoma to adenoma expression ratio, whereas PDGFRB and FZD2 were the most upregulated genes in the carcinoma samples and thus had the highest carcinoma to adenoma expression ratio (Figure 2). The average expression level of all the genes in the pathway did not change significantly for the adenoma and carcinoma samples (or stages), which centered around a ratio of 1. Analyzing the members at particular locations of the pathway suggested that different genes are differentially expressed and thus possibly differentially regulated, depending on the stage of the cancer tissues from which the expression data were obtained (Figure 2, Supplementary table 1). This demonstrated one of the findings that could be obtained with this integrative approach and would have been lost by analyzing all the pathway members. Analysis of all the pathway members as a whole, suggested no difference between the stages (i.e., ratio = 1) and did not identify a potential for any genes to be differentially regulated. FZD2 and FZD3 are different homologues of Frizzled, which is the receptor for Wnt signaling molecules and is known to be important in early development of colorectal cancer [8]. AXIN2 (Axin) is also a member of Wnt signaling and has been shown to be a mutation target in the development of colorectal adenoma [8]. Our analysis suggests that Wnt signaling members, such as different homologues of Frizzled receptor (FZD2, FZD3), may play a role in early (adenoma) and late (carcinoma) events of colorectal cancer progression. Wnt signaling is known to be activated during the earlier stages of colorectal cancer progression and is suggested to be involved in also the later stages of the progression [10]. Platelet-derived growth factor receptor beta (PDGFRB) is known to play a role in advanced and metastatic stages of colorectal cancer development [23]. It is noteworthy that most genes that showed differential expression in adenoma and carcinoma samples were receptors, i.e. Frizzled and PDGFRB. In support of this integrative pathway analysis, a previous study [21] also found that pathway information, i.e. whether it is a receptor, was important in identifying physiologically relevant functional groups. Currently, the drugs, Erbitux and Vectibix, used to treat colorectal cancer, target a receptor, EGFR (RTK) (Supplementary table 19). The location of the gene in the colorectal cancer pathway corresponds to the early cancer stage and the receptor region. The analysis appears to suggest that receptors could be important regulatory regions in colorectal cancer development, and as such, other receptors, i.e. FZD2, FZD3 and PDGFRB, could be possible candidates for drug development. However, more data is needed to confirm this relationship.

Next, we analyzed the possibility of dimensional (X and Y) distinction of the overall pathway, in other words, whether a group of pathway members defined by a stage (Y) or a cellular location (X) showed a significant difference in terms of expression levels, and the

presence of mutation and drug targets, as compared to the other members of the pathway. We set the X values to vary from 1 to 3, corresponding to the following locations: receptor/ligand (denoted as 1), cytosol (denoted as 2), and nucleus (denoted as 3). We examined the significance of grouping the expression profiles of the pathway members according to their X values. There was a significant grouping of both the adenoma and carcinoma expression profiles of the pathway members across the X groups (Table 1). In the adenoma and the carcinoma datasets, cytosolic members of the colorectal cancer pathway (X=2) had significantly lower gene expression values (Figure 3A, 3B, Table 2), on the other hand, nuclear members of the colorectal cancer pathway (X=3) had significantly higher gene expression values relative to other X groups (Figure 3C, 3D, Table 2).

Similar analysis was performed for the Y values, where the Y values ranged from 1 to 4, to correspond to the different stages, from normal tissue to early adenoma (denoted as 1), early to intermediate adenoma (denoted as 2), intermediate to advanced adenoma (denoted as 3), and advanced adenoma to carcinoma (denoted as 4). There was a significant difference across the Y groups for the adenoma but not the carcinoma datasets (Table 1), which may be attributed to a significant difference in the expression values of the pathway members with Y values of 1 as compared to the other pathway members (Y values of 2–4) in adenoma (Figure 3E, Table 2). In the adenoma tissue samples, the gene expression values of the pathway members, which play a role in the normal epithelium to early adenoma stage (Y=1) were expressed significantly higher than the other pathway members. The key genes that contribute to these results were BIRC5 (Survivin), FZD3, and AXIN2 (Axin), which are highly expressed and are located at Y=1, and PIK3CG, and PDGFRA, which are lowly expressed in the colorectal adenoma samples and located at Y=2 (Supplementary Table 1). This result is in line with our previous observation that a group of genes in a particular pathway could be differentially expressed from the other members in the pathway depending on the stage of the cancer. In addition to the adenoma tissue samples, we analyzed the expression values of the colorectal pathway members in the carcinoma tissue samples for possible distinctive patterns but found no significant grouping of the carcinoma expression with respect to Y (Figure 3F, Table 1, 2). If we consider that the colorectal cancer pathway includes the stages from normal epithelium to carcinoma, with several adenoma stages but only a single stage of carcinoma development, and also only one set of gene expression data from carcinoma tissues, there is likely insufficient information to distinguish the molecular events in the carcinoma pathway(see Figure 1).

3.2 Coherence of the colorectal cancer pathway

The KEGG cancer pathways represent a collective behavior of a group of proteins that underlies the disease, and as such, are built from proteins that belong to multiple signaling pathways. A coherence indicator has been defined as the ratio of the number of correlated gene pairs to the total number of gene pairs in a pathway, which is deemed significant based upon a statistical measure [20]. Using this indicator, the gene expression levels of the signaling and metabolic pathway members were shown to be coherent, suggesting that coherence may be an important measure of functionally related genes [20]. Cancer pathways, due to their very nature of involving multiple signaling groups, may not be expected to show the same level of coherence. Therefore, we examined whether the gene expression levels of the KEGG colorectal cancer pathway members were coherent, as compared to apoptosis, oxidative phosphorylation and proteasome pathway members, the latter were previously shown to be coherent [20]. The degree of coherence is determined by plotting the correlation in gene expression of the pathway members. We analyzed normal colorectal, adenoma and carcinoma samples. In normal colorectal tissue expression data, the oxidative phosphorylation and proteasome pathways show a distinct positive correlation among their pathway members, the apoptosis pathway show a slightly negative correlation,

whereas the correlation distribution of the colorectal cancer members are closer to the random distributions and hence uncorrelated (Figure 4A). In addition to the correlation distribution of the expression levels, we analyzed the cumulative distributions of the absolute values of the correlation coefficients. Similar results for the cumulative and non-cumulative correlation distributions are observed for normal colorectal tissue, i.e. the pathway is uncorrelated (Figure 4A, 4B). Oxidative phosphorylation and proteasome pathways are coherent in all 3 groups (normal, adenoma, and carcinoma tissues), while the colorectal cancer pathway is coherent only in the carcinoma samples (Figure 4B–4D). We compared whether the colorectal cancer and apoptosis pathways differed significantly from random at a correlation coefficient of 0.5 (Table 3). The colorectal cancer pathway appears to be coherent for colorectal carcinoma but not for the normal and adenoma datasets. This suggests that a cancer pathway may be coordinately regulated to achieve a biological function or phenotype, which, in this case, is the progression of the colorectal tissue to the tumor stage. The apoptosis pathway appears to be coherent in normal colorectal sample but not in the colorectal adenoma or carcinoma data (Table 3). This is in contrast to a previous report which used different gene expression data and found the data was coherent in the colorectal tumor but not in the normal samples [20]. In addition to the entire pathway, we studied the possibility of coherence of dimensional groupings in colorectal cancer and apoptosis pathways. Our analysis suggests that members within the pathways could be coherent, for example X=2 in the colorectal cancer pathway could be coherent in normal and adenoma samples (Supplementary table 20) even though the entire pathway may not be coherent (Table 3). However, more data is required to confirm these analyses.

In order to determine which proteins contribute to the coherence of the colorectal cancer pathway in the carcinoma stage, we obtained pairs of genes with an absolute correlation coefficient of at least 0.5 (Supplementary table 21). There were 683 correlated gene pairs in the carcinoma samples, most of which were specific to carcinoma (Figure 5A). On the other hand, the normal and adenoma samples had fewer correlated gene pairs and shared more than half of them with each other. The pairs of correlated genes specific to normal and adenoma tissues included mostly genes in the early stage members of the pathway, such as AKT homologues (Gene ID of 207 and 208), DVL homologues (Gene ID of 1856 and 1857), and FZD homologues (Gene ID of 8321, 8322, 8323, 8324). On the other hand, pairs of correlated genes specific to carcinoma included mostly pairs of genes from different parts of the pathway, such as TGFBR2 homologue (Gene ID of 91) from late stage of the pathway, APC (Gene ID of 324) from early stage of the pathway, KRAS (Gene ID of 3485) and MET (Gene ID of 4233) from mid stage of the pathway. (Supplementary Table 21). Next, we analyzed an absolute correlation coefficient of at least 0.8, suggesting these gene pairs are highly correlated, which are provided in Table 4 (0 indicates absolute correlation below 0.8 and 1 indicates absolute correlation of at least 0.8). We identified 10 gene pairs which were highly correlated in both the normal and adenoma but not in carcinoma samples (Figure 5B, Table 4). These pairs, unlike the carcinoma specific gene pairs, suggest coordination within the members of the Wnt signaling pathway, such as FZD, DVL and AXIN. In addition, these members are highly correlated with RAC, which is downstream of RAS oncogene (in the adenoma stage of the colorectal cancer pathway). The coordination between the members of the Wnt signaling pathway and RAC is supported by protein level interactions [10]. On the other hand, there were 38 gene pairs in the carcinoma samples, none of which were correlated in the normal or adenoma samples (Figure 5B, Table 4). These strongly correlated gene pairs may suggest direct or indirect protein-protein or transcriptional interactions specific to the carcinoma stage of the colorectal cancer pathway. For example, the list included the gene pair TGFBR1 and TGFBR2, whose protein products are known to interact directly and play a role in advanced stages of colorectal cancer [12]. The presence of a strong correlation between these two genes suggests that the interaction of these two receptors may be relevant only in the carcinoma stage, for the samples we

analyzed in this study. Another significantly correlated pair was MYC and MSH2 (Table 4). MYC is a transcriptional regulator of MSH2 [24]. Therefore, MYC driven regulation of MSH2 may be important in carcinoma but not in the normal or adenoma samples. In addition, BIRC5 was identified to be highly correlated with various other genes, such as ACVR1B and EGFR, suggesting that BIRC5 may also affect the regulation of ACVR1B and EGFR. While some of the strongly correlated gene pairs of the colorectal cancer pathway are supported by the literature, our analysis suggests there are others that may be correlated but currently are not supported by the literature. These results suggest potentially novel direct or indirect interactions, or common regulations by upstream molecules. Since the analysis identified the colorectal cancer pathway to be significantly coherent only in the carcinoma samples, further experimental investigation of these correlations is needed to confirm these novel regulatory mechanisms for colorectal carcinoma.

3.3 Mutation target analysis

In addition to gene expression, the distribution of mutation targets of the colorectal cancer pathway members was compared with respect to their X and Y values. The pathway members with high X values had higher frequencies of mutation targets. The nuclear members (X=3) had the highest number of mutation targets, followed by the cytosolic members (X=2) and the receptor/ligand members (X=1) (Table 5). Most of the mutation targets were concentrated at the later two stages of the pathway (Y=3, Y=4); covering the progression from intermediate adenoma to carcinoma, with the highest number of mutation targets found in the last stage, from late adenoma to carcinoma. In other words, colorectal cancer mutation targets were represented more in the late tumor stages of the KEGG pathway. This result is supported by the current view that mutations accumulate and increase as the tumor grows and cancer progresses [8]. Lastly, we analyzed whether the mutation targets were differentially expressed in the adenoma or carcinoma tissues as compared to the other pathway members, and found there was no statistically significant difference between the mutation targets and the other pathway members. This may be due to the sparsity of expression and mutation data. Note that our list of mutations targets is a collection of several genes which were reported to be mutated in at least a single sample. Therefore only some of these mutation targets may actually be mutated in the samples from which the expression data were collected. Furthermore, we do not know whether these mutations are the cause or the effect of the gene expression level changes. Currently, not enough mutation and gene expression data are available to perform an extensive mutation analysis with respect to the X and Y groups.

Mutations alter important residues or domains of proteins that could lead to alteration in the binding properties of proteins to other proteins or regulatory DNA sequences, thus the levels of other genes and proteins may change. Previous reports have suggested an association between mutation events and changes in the gene expression levels [25,26]. There is no evidence to suggest a correlation between the mutation of a gene and its own mRNA expression levels, however, a mutation in the TP53 gene has been shown to be correlated with its protein expression level in cancer [27]. This current study also did not find a strong relationship between a mutation target and its gene expression level for the colorectal cancer pathway members.

4 Concluding remarks

In this study, we analyzed expression and mutation data of the KEGG colorectal cancer pathway members. Previous studies focused predominantly on analyzing gene expression data for an entire signaling pathway and assessing whether the entire pathway is differentially expressed in a microarray dataset. Here we demonstrate an integrative analysis that investigates the distribution of expression values of the pathway members. Our analysis

incorporated the location of the pathway members and found the expression values and the number of mutation targets varied depending on the cellular location (X) and stage of the cancer (Y) (Figure 4, Tables 2, 3). Previous studies found signaling and metabolic pathways to be coherent, and we show that the colorectal cancer pathway is coherent as well, depending on the stage at which the expression data was obtained. The members of the KEGG colorectal cancer pathway showed some degree of correlation in the expression profile of the carcinoma data (GSE4183).

This analysis has the potential to help uncover the roles of the different genes in the pathways in the progression of colorectal cancer. We were able to analyze only the colorectal cancer pathway in this study. However, we anticipate that as more information on cancer pathways and expression data for the various cancer stages becomes available, this approach to pathway analysis could be more widely applicable and may help contribute to our understanding of the similarities and differences in the progression of cancer in the different tissues. Similarly, this integrative analysis could be applied to analyze the progression of other diseases or biological processes.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We thank Xuewei Wang for help with statistical analysis. This research was supported in part by Michigan State University (MSU) Quantitative Biology and Modeling Initiative Fellowship, the MSU Foundation, Michigan Economic Development Corporation, the National Science Foundation (BES 0425821), and the National Institutes of Health (1R01GM079688-01, 1R21CA126136-01 and 1R21RR024439-01). Authors thank Mao Tanabe from KEGG for his help in addressing the many questions we had about the KEGG database.

Abbreviations

AML	acute myeloid leukemia
APC	adenomatosis polyposis coli
CASP3	caspase 3
CML	chronic myeloid leukemia
Cytc	cytochrome c
DCC	deleted in colon cancer
ERK	extracellular signal regulated protein kinase
GSK-3β	glycogen synthase kinase-3-beta
hMLH1	human mutL homolog 1
KEGG	kyoto encyclopedia of genes and genomes
NCBI	national center for biotechnology information
PKB	protein kinase B
RTK	receptor tyrosine kinase
TCFLEF	T cell factor/lymphoid enhancer factor
TGFBR2	transforming growth factor beta receptor 2
VEGF	the receptor ligand vascular endothelial growth factor

References

1. Fearon ER, Vogelstein B. A genetic model for colorectal tumorigenesis. Cell. 1990; 61:759–767. [PubMed: 2188735]
2. Cardoso J, Boer J, Morreau H, Fodde R. Expression and genomic profiling of colorectal cancer. Biochim Biophys Acta. 2007; 1775:103–37. [PubMed: 17010523]
3. Edelman EJ, Guinney J, Chi JT, Febbo PG, Mukherjee S. Modeling Cancer Progression via Pathway Dependencies. PLoS Comput Biol. 2008; 4:e28. [PubMed: 18282083]
4. Kanehisa M, Goto S. KEGG: Kyoto Encyclopedia of Genes and Genomes. Nucleic Acids Res. 2000; 28:27–30. [PubMed: 10592173]
5. Kanehisa M, Goto S, Hattori M, Aoki-Kinoshita KF, et al. From genomics to chemical genomics: new developments in KEGG. Nucleic Acids Res. 2006; 34:D354–357. [PubMed: 16381885]
6. Kanehisa M, Araki M, Goto S, Hattori M, et al. KEGG for linking genomes to life and the environment. Nucleic Acids Res. 2008; 36:D480–D484. [PubMed: 18077471]
7. Dreesen O, Brivanlou AH. Signaling Pathways in Cancer and Embryonic Stem Cells. Stem Cell Rev. 2007; 3:7–17. [PubMed: 17873377]
8. Grady WM. Genomic instability and colon cancer. Cancer Metastasis Rev. 2004; 23:11–27. [PubMed: 15000146]
9. Söreide K, Janssen EA, Söiland H, Körner H, Baak JP. Microsatellite instability in colorectal cancer. Br J Surg. 2006; 93:395–406. [PubMed: 16555243]
10. Behrens J. The role of the Wnt signalling pathway in colorectal tumorigenesis. Biochem Soc Trans. 2005; 33:672–5. [PubMed: 16042571]
11. Mehlen P, Fearon ER. Role of the dependence receptor DCC in colorectal cancer pathogenesis. J Clin Oncol. 2004; 22:3420–8. [PubMed: 15310786]
12. Roman C, Saha D, Beauchamp R. TGF-beta and colorectal carcinogenesis. Microsc Res Tech. 2001; 52:450–7. [PubMed: 11170304]
13. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci USA. 2005; 102:15545–50. [PubMed: 16199517]
14. Li Z, Srivastava S, Findlan R, Chan C. Using dynamic gene module map analysis to identify targets that modulate free fatty acid induced cytotoxicity. Biotechnol Prog. 2008; 24:29–37. [PubMed: 18052188]
15. Li Z, Srivastava S, Yang X, Mittal S, et al. A hierarchical approach employing metabolic and gene expression profiles to identify the pathways that confer cytotoxicity in HepG2 cells. BMC Syst Biol. 2007; 11:1–21.
16. Breslin T, Krogh M, Peterson C, Troein C. Signal transduction pathway profiling of individual tumor samples. BMC Bioinformatics. 2005; 6:163. [PubMed: 15987529]
17. Mootha VK, Lindgren CM, Eriksson KF, Subramanian A, et al. PGC-1alpharesponsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. Nat Genet. 2003; 34:267–273. [PubMed: 12808457]
18. Jansen R, Greenbaum D, Gerstein M. Relating whole-genome expression data with protein-protein interactions. Genome Res. 2002; 12:37–46. [PubMed: 11779829]
19. Hannenhalli S, Levy S. Transcriptional regulation of protein complexes and biological pathways. Mamm Genome. 2003; 14:611–9. [PubMed: 14629111]
20. Yang HH, Hu Y, Buetow KH, Lee MP. A computational approach to measuring coherence of gene expression in pathways. Genomics. 2004; 84:211–7. [PubMed: 15203219]
21. Draghici S, Khatri P, Tarca AL, Amin K, et al. A systems biology approach for pathway level analysis. Genome Res. 2007; 17:1537–45. [PubMed: 17785539]
22. Futreal PA, Coin L, Marshall M, Down T, et al. A census of human cancer genes. Nat Rev Cancer. 2004; 4:177–83. [PubMed: 14993899]
23. Wehler TC, Frerichs K, Graf C, Drescher D, et al. PDGFRalpha/beta expression correlates with the metastatic behavior of human colorectal cancer: a possible rationale for a molecular targeting strategy. Oncol Rep. 2008; 19:697–704. [PubMed: 18288404]

24. Menssen A, Hermeking H. Characterization of the c-MYC-regulated transcriptome by SAGE: identification and analysis of c-MYC target genes. Proc Natl Acad Sci USA. 2002; 99:6274–9. [PubMed: 11983916]

25. Whyte DB, Holbeck SL. Correlation of PIK3Ca mutations with gene expression and drug sensitivity in NCI-60 cell lines. Biochem Biophys Res Commun. 2006; 340:469–75. [PubMed: 16376301]

26. Austinat M, Dunsch R, Wittekind C, Tannapfel A, et al. Correlation between beta-catenin mutations and expression of Wnt-signaling target genes in hepatocellular carcinoma. Mol Cancer. 2008; 7:21. [PubMed: 18282277]

27. Martinez-Delgado B, Robledo M, Arranz E, Infantes F, et al. Correlation between mutations in p53 gene and protein expression in human lymphomas. Am J Hematol. 1997; 55:1–8. [PubMed: 9136910]

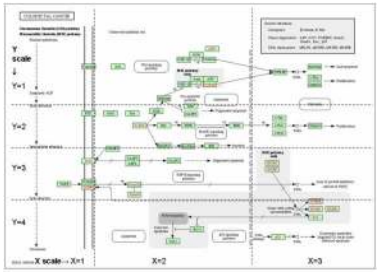


Figure 1.
Adapted from KEGG colorectal cancer pathway [20,21,22].

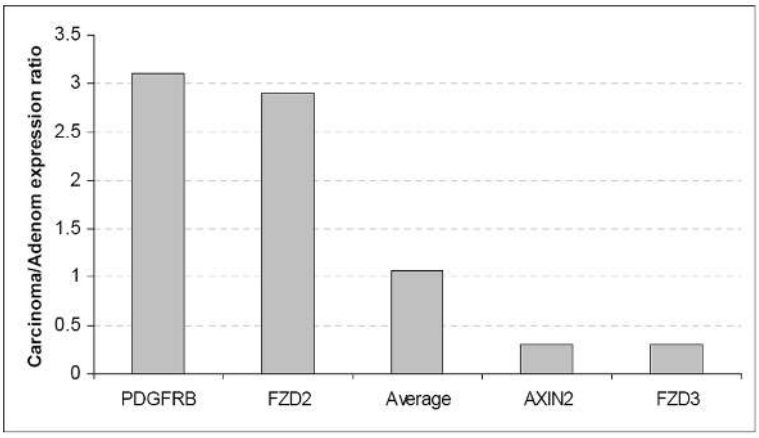
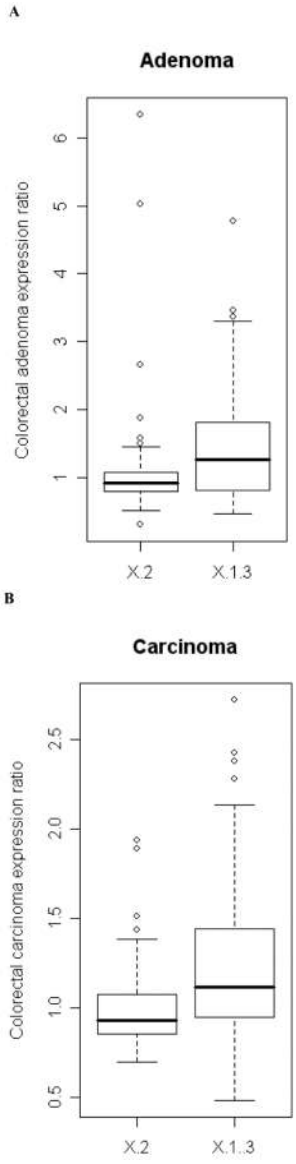
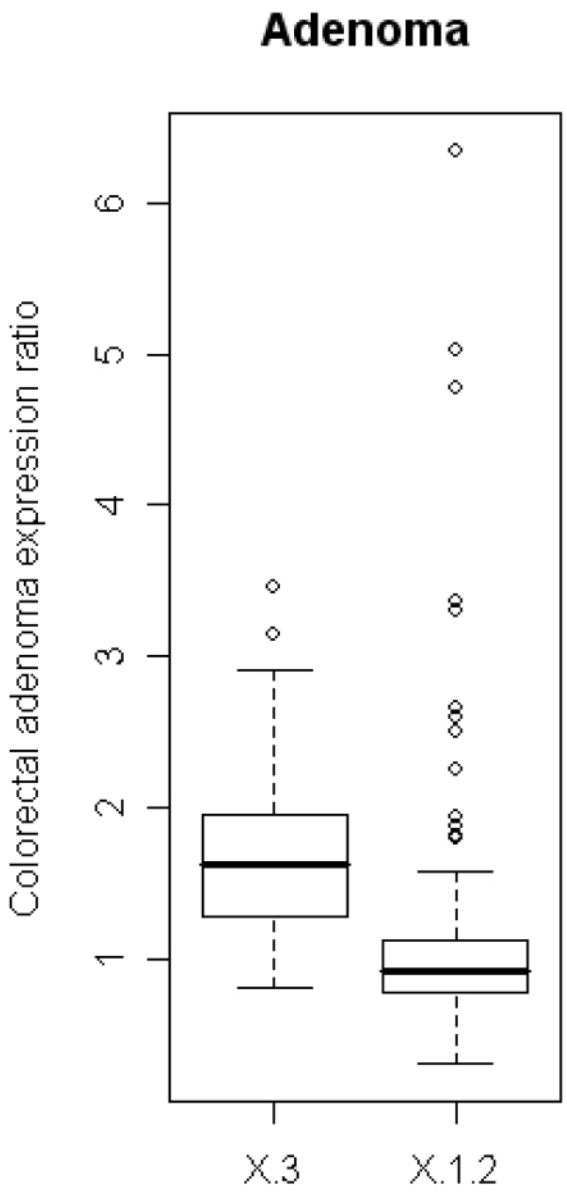


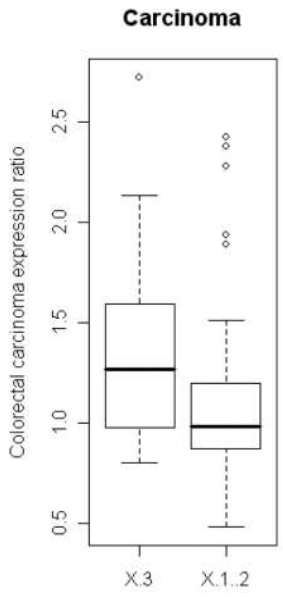
Figure 2.
The ratio of carcinoma expression ratio to the adenoma expression ratio given for the maximum and minimum 2 genes and the average.



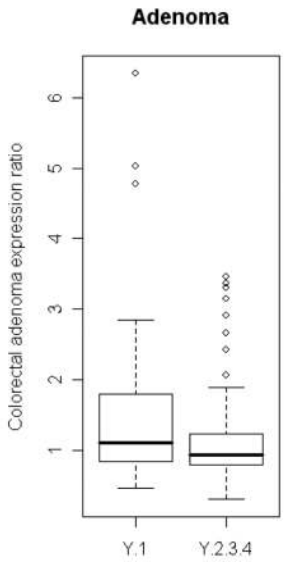
C



D



E



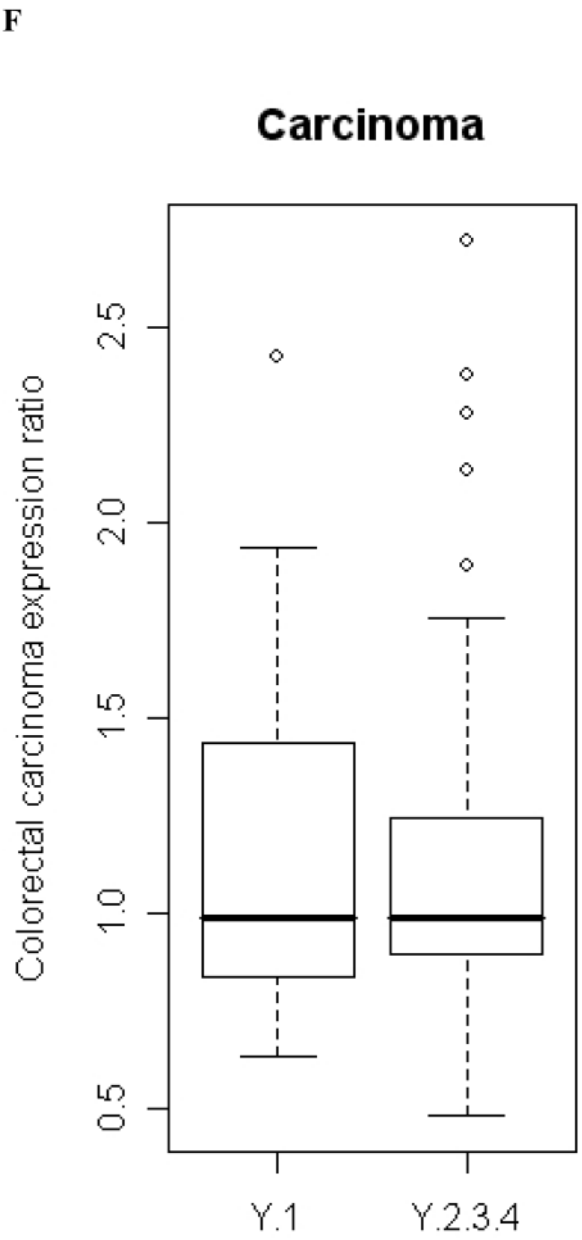
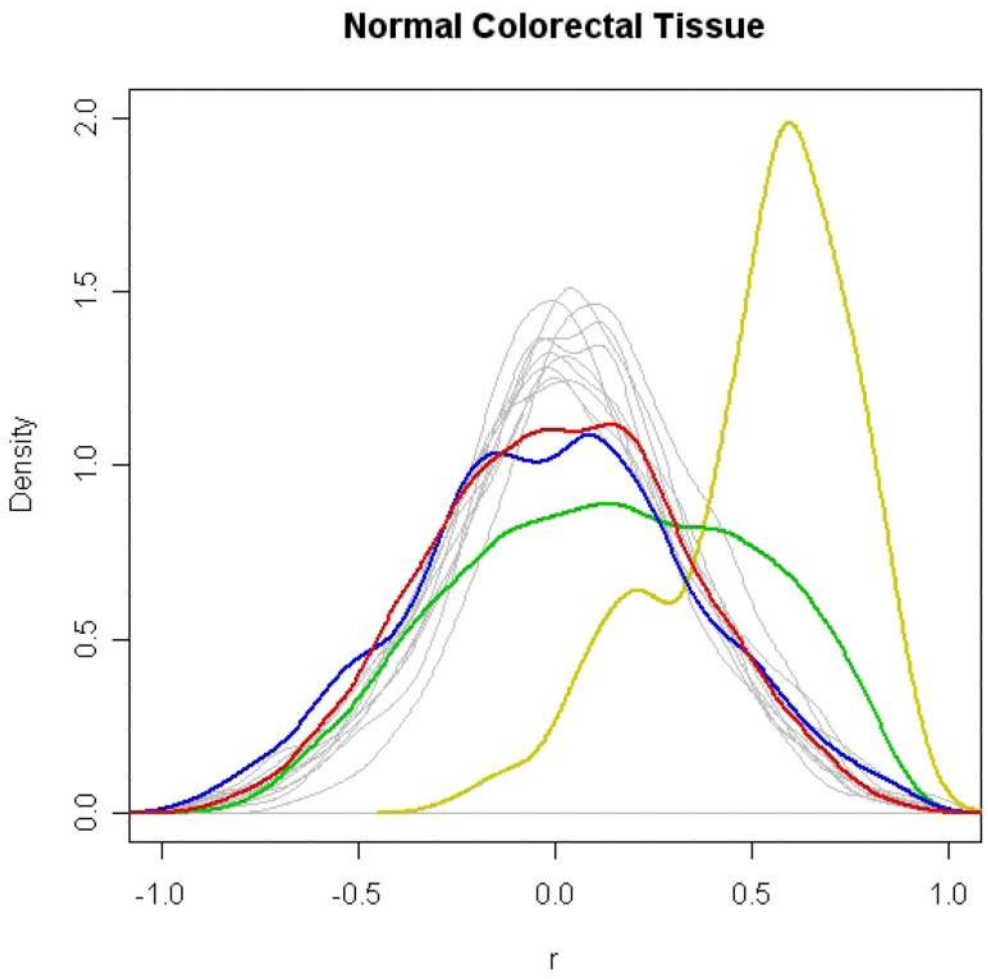
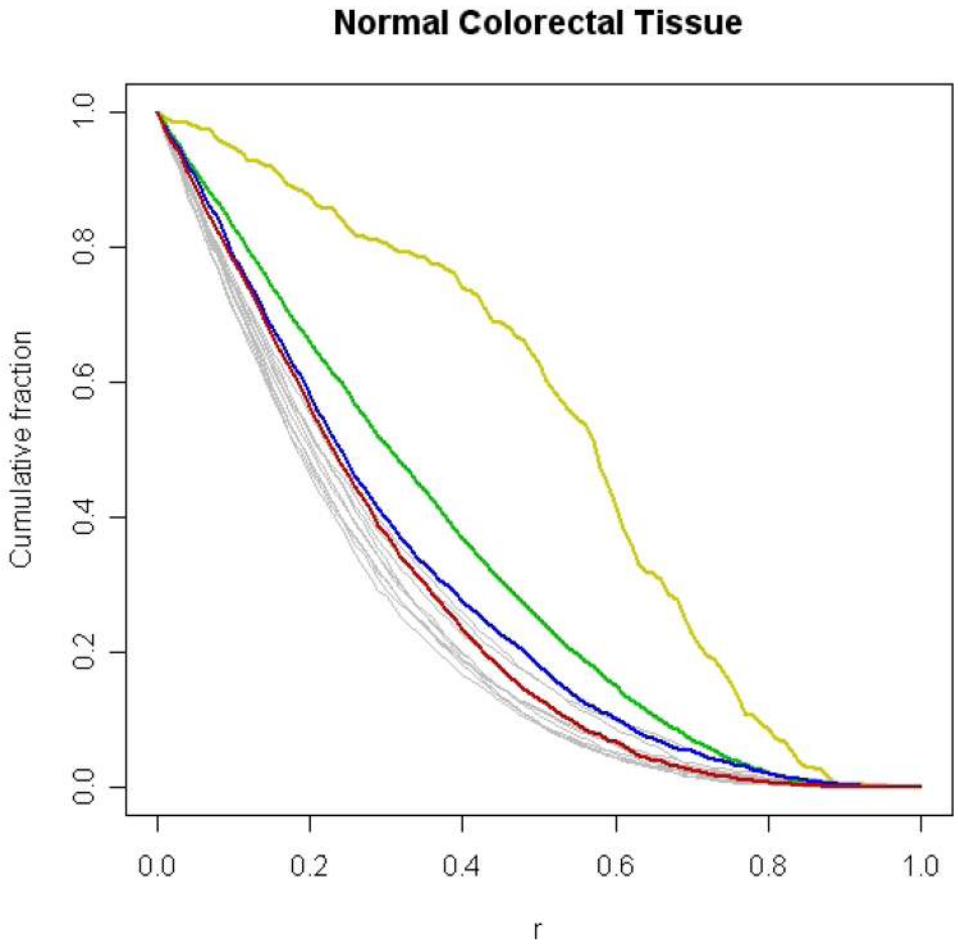


Figure 3.
X/Y dependent analysis of colorectal cancer pathway gene expression levels. **A)** Comparison of expression ratio values of X=2 to other X groups in adenoma. **B)** Comparison of expression ratio values of X=2 to other X groups in carcinoma. **C)** Comparison of expression ratio values of X=3 to other X groups in adenoma. **D)** Comparison of expression ratio values of X=3 to other X groups in carcinoma. **E)** Comparison of expression ratio values of Y=1 to other Y groups in adenoma. **F)** Comparison of expression ratio values of Y=1 to other Y groups in carcinoma.

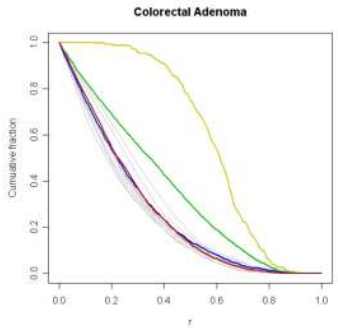
A



B



c



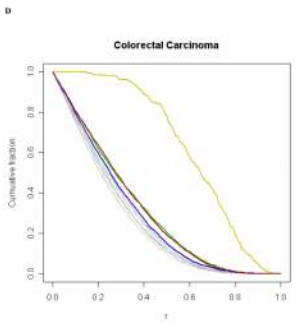


Figure 4.

Pathway correlation and cumulative fraction distributions. **A)** Pearson's correlation coefficient distributions among members of apoptosis, colorectal cancer, oxidative phosphorylation, and proteasome pathways shown together with 100 random selections of size 84 for GSE4183 expression data of normal colorectal tissue, **B–D)** Cumulative absolute value fractions of correlation coefficient distributions among members of apoptosis, colorectal cancer, oxidative phosphorylation, and proteasome pathways shown together with 10 random selections of size 84 for **B)** normal colorectal tissue, **C)** colorectal adenoma, and **D)** colorectal carcinoma. Random is shown as grey, colorectal cancer pathway is shown as red, apoptosis pathway is shown as blue, oxidative phosphorylation pathway is shown as green and proteasome pathway is shown as yellow.

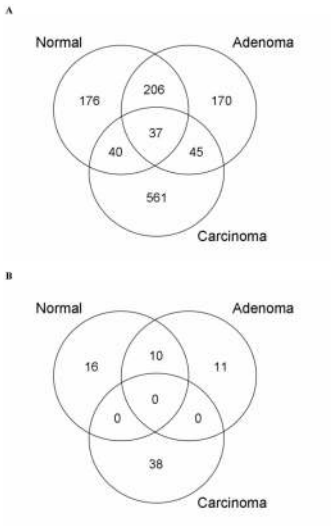


Figure 5.
Vend diagram of correlated pairs of genes in colorectal normal, adenoma and carcinoma samples. **A)** Correlated pairs with at least 0.5 absolute correlation coefficient. **B)** Correlated pairs with at least 0.8 absolute correlation coefficient.

Table 1

ANOVA p values

Group	Sample set	P value
X=1, 2, 3	Adenoma	0.0005
X=1, 2, 3	Carcinoma	0.0058
Y=1, 2, 3, 4	Adenoma	0.0265
Y=1, 2, 3, 4	Carcinoma	0.6136

Table 2

Pair-wise t- test p values

Group 1	Group 2	Adenoma p value	Carcinoma p value
X=1	X=2, 3	0.9470	0.2051
X=2	X= 1, 3	0.0038	0.0023
X=3	X=1, 2	0.0002	0.0149
Y=1	Y=2, 3, 4	0.0049	0.8826
Y=2	Y=1, 3, 4	0.1654	0.3116
Y=3	Y=1, 2, 4	0.0882	0.3050
Y=4	Y=1, 2, 3	0.7675	0.5209

Table 3

Apoptosis and colorectal cancer pathway coherence at correlation coefficient of 0.5

Sample set	Apoptosis pathway	Colorectal cancer pathway
Normal	0.021	0.255
Adenoma	0.182	0.244
Carcinoma	0.119	0.002

Proteomics Clin Appl. Author manuscript; available in PMC 2011 February 4.

Table 4

Correlated pairs (with at absolute correlation coefficient level of 0.8 as the threshold) in the colorectal cancer pathway for normal, adenoma and carcinoma samples

Gene ID 1	Gene Name 1	Gene ID 2	Gene Name 2	Normal	Adenoma	Carcinoma
91	ACVR1B	1956	EGFR	0	0	1
208	AKT2	4087	SMAD2	0	0	1
324	APC	5602	MAPK10	0	0	1
332	BIRC5	91	ACVR1B	0	0	1
332	BIRC5	324	APC	0	0	1
332	BIRC5	1956	EGFR	0	0	1
332	BIRC5	2956	MSH6	0	0	1
332	BIRC5	4436	MSH2	0	0	1
332	BIRC5	5602	MAPK10	0	0	1
332	BIRC5	6932	TCF7	0	0	1
332	BIRC5	8313	AXIN2	0	0	1
842	CASP9	6934	TCF7L2	0	0	1
2353	FOS	7046	TGFBR1	0	0	1
2956	MSH6	1630	DCC	0	0	1
2956	MSH6	8313	AXIN2	0	0	1
3845	KRAS	1630	DCC	0	0	1
3845	KRAS	2956	MSH6	0	0	1
3845	KRAS	6934	TCF7L2	0	0	1
4233	MET	5604	MAP2K1	0	0	1
4609	MYC	332	BIRC5	0	0	1
4609	MYC	4436	MSH2	0	0	1
4609	MYC	6932	TCF7	0	0	1
5880	RAC2	5293	PIK3CD	0	0	1
6654	SOS1	4436	MSH2	0	0	1
6654	SOS1	8322	FZD4	0	0	1
6655	SOS2	324	APC	0	0	1
7040	TGFB1	5159	PDGFRB	0	0	1
7043	TGFB3	23533	PIK3R5	0	0	1

Gene ID 1	Gene Name 1	Gene ID 2	Gene Name 2	Normal	Adenoma	Carcinoma
7976	FZD3	5881	RAC3	0	0	1
8313	AXIN2	6932	TCF7	0	0	1
8322	FZD4	4436	MSH2	0	0	1
8326	FZD9	4087	SMAD2	0	0	1
83439	TCF7L1	332	BIRC5	0	0	1
83439	TCF7L1	4436	MSH2	0	0	1
83439	TCF7L1	5602	MAPK10	0	0	1
130399	ACVR1C	91	ACVR1B	0	0	1
130399	ACVR1C	1956	EGFR	0	0	1
130399	ACVR1C	5295	PIK3R1	0	0	1
1857	DVL3	5159	PDGFRB	0	1	0
1857	DVL3	10000	AKT3	0	1	0
5296	PIK3R2	23533	PIK3R5	0	1	0
5879	RAC1	1857	DVL3	0	1	0
5879	RAC1	5881	RAC3	0	1	0
5879	RAC1	8321	FZD1	0	1	0
5879	RAC1	8324	FZD7	0	1	0
5881	RAC3	23533	PIK3R5	0	1	0
7040	TGFB1	1630	DCC	0	1	0
8313	AXIN2	5159	PDGFRB	0	1	0
8323	FZD6	10000	AKT3	0	1	0
207	AKT1	1857	DVL3	1	0	0
207	AKT1	5881	RAC3	1	0	0
208	AKT2	207	AKT1	1	0	0
208	AKT2	1857	DVL4	1	0	0
208	AKT2	5291	PIK3CB	1	0	0
208	AKT2	5881	RAC3	1	0	0
1857	DVL3	5291	PIK3CB	1	0	0
1857	DVL3	8321	FZD1	1	0	0
7157	TP53	8323	FZD6	1	0	0
7157	TP53	8324	FZD7	1	0	0

Gene ID 1	Gene Name 1	Gene ID 2	Gene Name 2	Normal	Adenoma	Carcinoma
8313	AXIN2	207	AKT1	1	0	0
8313	AXIN2	208	AKT2	1	0	0
8313	AXIN2	7157	TP53	1	0	0
8313	AXIN2	8321	FZD1	1	0	0
8313	AXIN2	8323	FZD6	1	0	0
8313	AXIN2	8324	FZD7	1	0	0
1857	DVL3	5881	RAC3	1	1	0
1857	DVL3	8323	FZD6	1	1	0
7157	TP53	5881	RAC3	1	1	0
8313	AXIN2	1857	DVL5	1	1	0
8313	AXIN2	5881	RAC3	1	1	0
8321	FZD1	5881	RAC3	1	1	0
8321	FZD1	8324	FZD7	1	1	0
8323	FZD6	5881	RAC3	1	1	0
8323	FZD6	8324	FZD7	1	1	0
8324	FZD7	5881	RAC3	1	1	0

Table 5

Mutation frequency values of X/Y groups

Group	Average mutation frequency
X=1	0.0455
X=2	0.1702
X=3	0.3333
Y=1	0.0833
Y=2	0.0789
Y=3	0.3333
Y=4	0.5714