Overlapping Gene Expression in Fetal Mouse Intestine Development and Human Colorectal Cancer

Michael Hu[1] and Ramesh A. Shivdasani[1,2]

[1]Dana-Farber Cancer Institute and [2]Department of Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, Massachusetts

Abstract

Pathways relevant to cancer are well known to overlap with fetal development, as reflected in reactivation of embryonic genes in tumors. However, molecular evidence for this notion has gathered in piecemeal fashion, and systematic approaches have rarely been applied to gauge the extent and global characteristics of the overlap in gene expression between developing tissues and cancer. The fraction of genes that is expressed aberrantly in a given cancer and also developmental in primary function is unknown, and the tissue specificity of recapitulated gene expression remains unexplored. We developed a statistical method to relate expression profiles from human colon cancer and diverse nonintestinal tumors to transcripts that decline in expression with epithelial differentiation in the fetal mouse gut. For genes that are overexpressed in colon cancer, we computed 8% to 19% likelihood that they were expressed transiently during epithelial morphogenesis in intestine development. Among genes dysregulated in other tumors, the corresponding likelihood fell between 1% and 6%. Similarly, low probabilities were obtained when we compared genes not overexpressed in colon cancer with transcriptional profiles in intestine organogenesis. Genes that increase after fetal gut epithelial differentiation were not differentially represented between cancerous and normal colon. Our findings systematically characterize the global extent and tissue specificity of developmental expression programs in colorectal cancer and illustrate the use of such an approach to identify candidate biomarkers and therapeutic targets. (Cancer Res 2005; 65(19): 8715-22)

Introduction

Molecular mechanisms of embryonic development are recognized to correlate with those in cancer, and a growing body of evidence highlights various signaling, transcriptional, and metabolic pathways that are shared between organogenesis and malignant tumors (1–3). Additionally, important properties of tumors including tissue invasion, viability at distant sites, and drug resistance, correlate strongly with the degree of histologic differentiation in resected specimens. Such considerations have popularized the idea that tumor cells represent reversion to a primitive state, although definition of such states is imprecise in both concept and molecular characterization. The supporting evidence is largely anecdotal, and systematic approaches have rarely been applied to gauge the extent and global characteristics of the overlap in gene expression between developing tissues and cancer. Questions such as what fraction of the genes expressed aberrantly in a given cancer reflect reactivation of a developmental program, or whether recapitulated gene expression is characteristic to the affected tissue, remain unanswered. Besides their relevance to tumor biology, these questions have practical implications. First, oncofetal markers can serve as useful tools in cancer diagnosis and in monitoring response to therapy (4, 5). Second, cancer treatments often are limited by severe toxicity that reflects expression of the drug target in unaffected tissues. If tumors depend on the expression of some proteins that are absent in adult tissues, then as therapeutic targets, such proteins might confer a wide therapeutic window.

Computational analysis of suitable mRNA expression data sets could potentially yield systematic approximations of the true underlying statistics on overlapping gene expression between developing and cancerous tissues. One such study (6) interrogated microarray-based gene expression profiles across human medulloblastomas and mouse cerebellar development. The authors used a pattern classification (singular value decomposition) approach to reveal that transcripts increased in expression in human medulloblastoma significantly reflect early mouse cerebellar development, whereas transcripts reduced in this disease correspond to a later, complementary program of gene expression.

The intestinal mucosa is organized into crypts, which house replicating progenitor cells, and villous projections lined by postmitotic epithelial cells with differentiated morphology and functions (7, 8). In a significant developmental transition, the gut endoderm first acquires villous character between 13 and 15 days in mouse gestation. We have assembled over 65,000 serial analysis of gene expression (SAGE; ref. 9) tags (representing over 10,000 unique mRNAs) from the mouse small intestine at 12, 13, and 15 days post-coitus. The complete data set, found at http://genome.dfci.harvard.edu/GutSAGE, shows nearly twice as many significant changes in gene expression between 13 and 15 days post-coitus than between 12 and 13 days post-coitus, revealing notable modulation of gene activity in conjunction with epithelial histogenesis. Genes with reduced expression following the villus transition may serve biological functions that are confined to the period of organogenesis. We recently showed that many such transcripts are absent from the adult gut, and for a small group of developmentally repressed genes, we presented experimental evidence that up to one fifth of the transcripts may be reexpressed in human colorectal tumors (10). Here, we extend these limited findings by applying a computational strategy toward a larger number of transcripts.

We mapped all existing human colon cancer gene expression profiles and a diverse set of expression profiles from nonintestinal tumors onto our data set of mouse intestine development. We applied a conditional probability method to approximate the

Note: Supplementary data for this article are available at Cancer Research Online (http://cancerres.aacrjournals.org/).

Requests for reprints: Ramesh A. Shivdasani, Dana-Farber Cancer Institute, 44 Binney Street, Boston, MA 02115. Phone: 617-632-5746; Fax: 617-632-4471; E-mail: ramesh_shivdasani@dfci.harvard.edu.

doi:10.1158/0008-5472.CAN-05-0700

Table 1. Sizes of CAN_{INT} and NC_{INT} gene expression sets as they were mapped from human across to homologous mouse genes represented in SAGE profiles of mouse gut development

Colon cancer resource	Filter 1 result (no. unique transcripts)		Filter 2 result (no. mouse homologues)		Filter 3 result (no. with SAGE entry)	
	CAN	*NC*	CAN_{HOMOL}	NC_{HOMOL}	CAN_{INT}	NC_{INT}
Microarray data						
mArray_adCa1	160	1,340	69	481	51	49
mArray_adCa2	150		68		52	40
mArray_adCa3	147		62		46	48
mArray_adCa4	157		70		51	38
mArray_adCa5	176		80		49	45
mArray_GCM_colon	347	3,018	178	1,310	128	105
SAGE libraries (tag no.)						
NC1 versus Tu_98 (14,300)	141	619	65	305	42	89
NC2 versus Tu_102 (7,400)	69	345	30	169	18	49

NOTE: Detailed methods are described in Materials and Methods.

degree to which human colon cancers show tissue-specific recapitulation of developmentally regulated genes. For genes that are overexpressed in colon cancer, our statistical analysis yields 8% to 19% likelihood that they were expressed transiently during gut epithelial morphogenesis in development. Among genes overexpressed in other malignancies, the corresponding probability falls between 1% and 6%. Our findings systematically estimate the extent to which cancer gene dysregulation coincides with the developmental expression program and show the tissue specificity of this process.

Materials and Methods

Treatment of human colon cancer data sets. To capture a comprehensive collection of matched tumor and normal tissue expression data, we queried Oncomine (http://141.214.6.50/oncomine/main/index.jsp) for all results on matched, well-characterized colon cancer compared with normal tissue (11). This search yielded two adequate microarray and two SAGE expression data sets. Starting with SAGE libraries (12) NC_1 versus Tu_98 (14,300 unique SAGE tags) and NC_2 versus Tu_102 (7,400 tags), we first mapped every SAGE tag against all nonredundant human Unigene entries (http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=unigene) and isolated transcripts (*CAN*) showing >2-fold increase in tumor compared with normal tissue ($P < 0.005$) or a control set (*NC*) for which tumor/normal tissue expression ratios fall between 0.8 and 1.2 ($P > 0.6$). A second filter applied HomoloGene (http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=homologene) criteria and retained only those human genes with a reliable murine homologue. Finally, consideration was limited to transcripts with at least one tag entry in our fetal intestine SAGE database (http://genome.dfci.harvard.edu/GutSAGE), and genes were removed arbitrarily from the *NC* set with the dual purpose of bringing the two final sets (CAN_{INT} and NC_{INT}) to a comparable size and to introduce additional randomness. For both the SAGE and following microarray data sets, Table 1 lists transcript numbers after application of the filters used to obtain CAN_{INT}.

The first microarray study (13, 14) evaluated carcinomas and normal colon tissue from 18 patients for expression corresponding to ~3,200 full-length human cDNAs and 3,400 expressed sequence tags (yielding ~6,500-gene coverage in aggregate). Because this limited interrogation could potentially lead to biased correlations, we applied statistical bootstrapping to generate multiple subsets from partially overlapping sample groups and tested each expression subset independently. We randomly bootstrapped five subsets (*mArray_adCa1* through *mArray_adCa5*), each derived from six tumor samples chosen randomly from the group of 18. For each subset, we selected transcripts showing >2-fold higher mean expression ($P < 0.001$) in tumors over normal tissue across all six sample pairs. Transcripts were then mapped to a nonredundant set of human LocusLink entries (http://www.ncbi.nlm.nih.gov/projects/LocusLink) followed by mapping all human LocusLink entries to their mouse homologues using the HomoloGene database. The corresponding mouse genes constitute a set designated CAN_{HOMOL}. CAN_{INT} contains all such genes with a record in our gut development SAGE database. For the control data set not related to cancer, we scanned the 18 samples for 1,340 transcripts whose mean expression in every case is similar in the tumor and normal tissue (ratio between 0.8 and 1.2, $P > 0.6$). The mouse homologues of 481 such genes (NC_{HOMOL}) were partitioned into five random gene sets, which we paired with the five CAN_{INT} groups (Table 1).

The second microarray study (15) probed adenocarcinomas and normal tissue from three patients for expression corresponding to ~14,000 Genbank accession nos. (Affymetrix Hu6800 and Hu35KsubA gene chips). Unlike the above example, the data are available on a raw expression scale, without *P* for comparisons. We therefore set a more stringent variable and isolated transcripts showing >2.5-fold mean expression in tumors over the matched normal tissue across all sample pairs; the resulting gene set is denoted *mArray_GCM_colon*. Filters similar to the above treatment were used to arrive at the corresponding sets CAN_{HOMOL} and CAN_{INT}. To generate the control, noncancer data set NC_{INT}, we scanned the paired samples for transcripts with similar mean expression in tumor and normal tissues (ratio between 0.9 and 1.1) and randomly selected 200 of 1,300 such transcripts to bring the noncancer and cancer sets to comparable size.

Treatment of data sets from human nonintestinal cancers. The principle of applying successive filters is similar to that described above for colon cancer expression data sets. We first extracted all transcripts that are overexpressed in tumors compared with matched normal tissue. The actual variables differ slightly for each data set, as described below. Where available (liver and plasma cell), we accepted the variables for differential expression from the original reports; for the remaining three data sets (prostate and breast), we chose expression cutoffs and *P*s that are common across the microarray literature and also returned a data set sufficiently large for statistical analysis. Owing to the heterogeneity of these large data sets and our demand of differential expression across >15 or >10 profiled tumors, we chose 1.5- and 1.75-fold cutoffs for the prostate and breast data, respectively. The numerical readouts in microarray analysis are well known

to underestimate, often significantly, the differential expression between cancer and normal tissues (15, 16), and higher arbitrary cutoffs yielded too few genes for further analysis; to reduce chance events, we avoided demanding altered gene expression in fewer tumors.

A second filter eliminated genes that lack a reliable mouse homologue, as specified in LocusLink and annotated in HomoloGene. The final filter retained only genes with a reliable entry in our mouse fetal intestine SAGE database. The resulting groups are denoted as nonintestinal cancer (*NIC*) gene sets, and Table 2 lists the sizes of the intermediate sets resulting from application of each filter.

Breast cancer. We used data from a study in which microarrays containing probes for 12,000 genes were interrogated with RNA derived from 12 invasive ductal carcinomas and three normal breast tissue samples (17). In the first filter, we considered transcripts that are elevated in >10 tumors with mean tumor/normal expression ratio of >1.75.

Liver tumors. Data were extracted from a study that characterized mRNA expression using ~17,400-gene chips on 82 hepatocellular carcinoma and 74 nontumorous liver samples (18). For the first filter, we took the reported list of ~1,640 most differentially expressed transcripts between tumor and normal samples (Bonferroni corrected $P < 0.01$; see Supplementary Section of the original report; ref. 18) and considered genes with a mean tumor/normal expression ratio of >2.

Prostate cancer. One study (Prostate1) compared gene expression from 52 adenocarcinomas and 50 normal samples using microarrays with probes for ~12,600 genes (19). Owing to heterogeneity of samples in the large cohort, we took transcripts with a "Present" call (expressed at detectable levels) in >15 of the 52 tumor samples and mean tumor/normal expression ratio of >1.5. For the data (Prostate2) from paired SAGE libraries (http://cgap.nci.nih.gov/SAGE), constructed from a microdissected tumor and adjacent normal epithelium, we required a ratio of tumor/normal SAGE tags of >1.5 ($P < 0.01$).

Plasma cell tumors. The data derive from a study that used Affymetrix microarrays to compare gene expression profiles from nine patients with multiple myeloma and eight myeloma cell lines with those of nonmalignant plasma cells (20). In our first filter, we took the list of 250 genes that the authors reported as being significantly overexpressed in malignant plasma cells (>2-fold, $P < 0.05$).

Results

Statement of the problem and data set resources. The main question and approach are depicted in Fig. 1. We define the rate of developmental gene recapitulation in cancer as the likelihood that a given gene that is activated in colon cancer was expressed at the highest level during gut epithelial morphogenesis. This is cast as the conditional probability Pr(DEV|CAN), or the probability of a given gene belonging to the developmental program, conditioned on its overexpression in cancer. The unknown variable, the true likelihood of recapitulation, is approximated by empirically determining the underlying Pr(DEV|CAN) (the ratio of the size of the subset x to that of the set CAN in Fig. 1). This value is computed for multiple sets of genes (CAN) that are overexpressed in colon cancer, derived from different experiments and profiling methods. We can thus empirically arrive at an approximation for Pr(DEV|-CAN) on each set, resulting in a distribution of likelihood values (right curve in Fig. 1). A similar approximation of developmental gene recapitulation in normal (noncancer, NC) tissues gives the distribution of the complementary probabilities Pr(DEV|NC) (Fig. 1, left curve). If the cancer and noncancer gene sets are of roughly equal size and the difference between the two empirical distributions (Fig. 1) is statistically significant, one would conclude that developmental gene expression programs are recapitulated at a higher rate in cancer than in the normal tissue. Note that our definitions do not say that more cancer genes are of developmental origin than non cancer genes, nor that more genes overexpressed in cancer are of developmental significance than not.

Table 2. Resulting sizes of *NIC* gene sets as they were mapped from human to homologous mouse genes represented in SAGE profiles of mouse intestine development

Tumor type	Filter 1 (no. human cancer genes)	Filter 2 (no. mouse homologs)	Filter 3, *NIC* (no. represented in developing gut)
Breast	187	168	115
Liver	102	90	57
Plasma cell	157	128	89
Prostate1	48	37	22
Prostate2	87	62	49

NOTE: Detailed methods are described in Materials and Methods.

We considered all transcripts in our mouse intestinal SAGE data set with significantly lower ($P < 0.01$) expression at E15 than at earlier stages (Fig. 2*A*) and denote the set of 254 transcripts that satisfy this rigorous criterion as DEV_{INT} set 1. Next, we obtained all public expression profiles that compare normal and malignant colorectal tissue by microarray or SAGE as described in Materials and Methods. These cancer data derive from human samples, whereas the developmental data profile mouse tissues; accordingly, the first task is to match human and mouse transcripts. We considered only those human genes with a mouse homologue, as specified through the HomoloGene database.

Statistical analysis of overlapping gene expression. As described in Materials and Methods and summarized in Table 1, we derived sets of genes that are overexpressed in colon cancer and whose mouse homologue is represented in SAGE profiles of the developing gut. The sets designated CAN_{INT} contain transcripts showing >2- or >2.5-fold increase in tumors compared with normal tissue. For one of the two microarray studies (13, 14), we generated five randomly bootstrapped sets (*mArray_adCa1* through *mArray_adCa5*), each containing transcripts with >2-fold higher mean expression ($P < 0.001$) in tumors over the normal tissue. For both SAGE and microarray cancer data, the control NC_{INT} sets represent genes whose mean expression was not significantly different between tumor and normal tissues ($P > 0.6$). We thus generated eight pairs of human gene sets that are overexpressed in colorectal cancer (CAN_{INT}) or nearly equally expressed in cancerous and normal intestine (NC_{INT}).

To analyze these expression data sets, we adopted the following conditional statistics formulations:

$$\Pr(x \in DEV_{INT} \mid x \in CAN_{INT}) \quad \text{(A)}$$

$$\Pr(x \in DEV_{INT} \mid x \in NC_{INT}) \quad \text{(B)}$$

Eq. A is the conditional probability that any gene x, given that it is overexpressed in colon cancer, is also repressed after E12 or E13 in mouse gut development. The second conditional is the distribution on the probability of any gene x, given that it is not overexpressed in intestinal cancer, being repressed in gut development. It provides

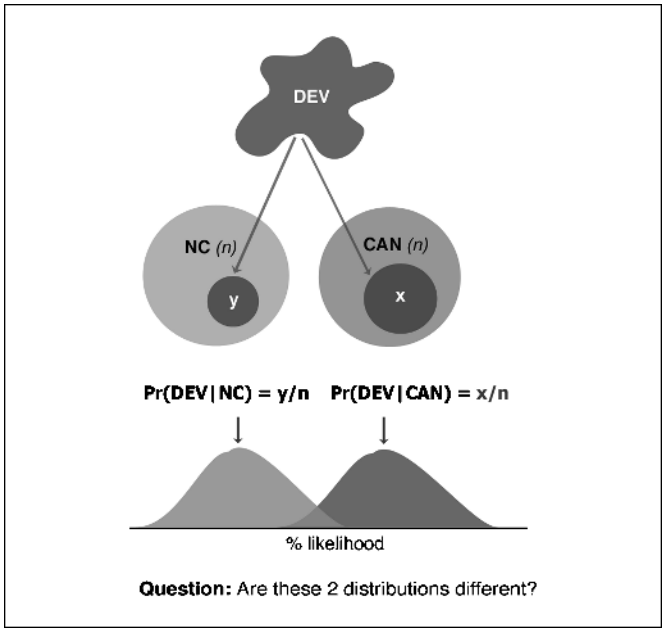


Figure 1. Diagrammatic representation of recapitulation of developmental genes in cancer and of the principal question posed in this study. The likelihood that a gene overexpressed in colon cancer was highly expressed during development of the gut is cast as the conditional probability Pr(DEV|CAN). For equal sizes (*n*) of cancer (CAN) and noncancer (NC) gene sets, we empirically approximated the underlying Pr(DEV|CAN) (representing the ratio x/n) on multiple cancer gene sets and Pr(DEV|NC) in the normal tissue counterpart. This analysis returns two independent distributions of likelihood values. If the difference between the two empirical distributions is statistically significant, it would indicate that the rate of recapitulation of a developmental program is greater in cancer than in the corresponding noncancerous tissue.

the control necessary to accept or reject the hypothesis that the two conditionals (Eqs. A and B) represent different underlying distributions (Fig. 1).

We computed the conditional probabilities in Eqs. A and B using the eight pairs of CAN_{INT} and NC_{INT} gene expression data sets (Table 1) and determined that the distribution on the likelihood of any gene being developmentally regulated in the intestine, given that it is overexpressed in colon cancer, falls between 7.7% and 18.8% (x/n_1 in Fig. 2*B* and *C*). However, if a gene is not overexpressed in colon tumors, its likelihood of being repressed during gut development (y/n_2 in Fig. 2*B* and *C*) is at most 6.7% and typically much lower. An unpaired, nonparametric (Mann-Whitney) *t* test on these data reveals the difference in distribution between the two conditional probabilities to be statistically significant ($P < 0.0006$). Indeed, the estimated probability of a given gene being silenced during intestine development is, on average, three to four times greater if it is overexpressed in colon cancer than if it is not.

To reduce spurious correlations resulting from data biases, we perturbed our calculations by using a larger developmental gene collection, DEV_{INT} set 2, with relaxed statistical criteria. This expanded set contains 451 transcripts with lower intestine expression at E15 compared with earlier stages ($P < 0.025$) and the potential addition of noise creates adversity for the hypothesis under investigation. When DEV_{INT} set 2 was used to approximate the conditional probabilities given by Eqs. A and B, absolute percentage values differed slightly but the general trends and statistical significance of the comparisons remained similar (Fig. 2*C*). This similarity in the face of adversity implies that the different distributions on Eqs. A and B represent true underlying biological differences. The results thus suggest a measurable bias for developmentally repressed genes to be reactivated in tumors of the same origin and they support a conclusion we recently reported on a smaller, experimental scale (10). Notably, gene expression in human colorectal cancer mimics only a portion of the transcriptional program that is active during intestine organogenesis.

Assessment of tissue specificity. The apparent reactivation of embryonic gene expression programs is not necessarily specific to the tissue of tumor origin. Gene overexpression in tumors could represent two possibilities. In one case, genes silenced during gut development could reactivate selectively in intestinal tumors, reflecting reversal of tissue-specific epigenetic regulatory mechanisms. Alternatively, reactivation of developmental genes in cancer might represent a non–tissue-specific process, where genes repressed during gut development are appropriated for malignant behaviors in diverse cell types. These scenarios differ both in their implications and in probable underlying mechanisms. Tissue-specific gene reactivation implies that oncogenesis represents some degree of developmental reversal, whereas non–tissue-specific reactivation points simply to shared molecular features between development and cancer. Of note, these implications apply only to the consideration of groups of genes; overexpression of any single gene in fetal and tumor tissues may reflect either coincidence or a cellular process that is common to both processes.

To assess the tissue specificity of developmental gene reexpression, we applied a third statistical formulation and evaluated gene expression data sets from *NIC*:

$$\Pr(x \in DEV_{INT} \mid x \in NIC) \qquad \text{(C)}$$

This conditional is the distribution on the probability that a gene *x*, given that it is activated in some nonintestinal cancer, is also repressed after E12 or E13 in mouse gut development. It assesses the specificity with which gut developmental gene expression is recapitulated in colon cancer, and we proposed that the conditionals given by Eqs. A and C form different distributions. The conditional probabilities were computed by taking the ratio of the number of genes found in the intersection of the appropriate gene sets, as shown in Fig. 3. To compute values for breast cancer, for example, we take the ratio y/n (Fig. 3*A*), where y is the number of genes in the intersection of the DEV_{INT} and breast cancer gene sets, and n is the number of genes in the breast *NIC* set with corresponding entries in fetal gut SAGE libraries (Table 2).

We extracted tumor expression profiles on various nonintestinal human cancers from public data sets. To avoid biases that may result from reliance on a single gene expression platform, we used results from both SAGE and microarray studies on diverse tumors. For five independent *NIC* gene sets, we computed the conditional probabilities given by Eq. C to fall between 1.3% and 5.6% (y/n in Fig. 3*B, column 1*). Thus, the estimated probability of a gene being developmentally silenced in the intestine is, on average, 2.5 to 3 times greater if it is overexpressed in colon cancer than if it is dysregulated in other tumors. An unpaired, nonparametric *t* test between the eight data points for Eq. A and five data points for Eq. C indicates different underlying distributions ($P < 0.001$), and there was no appreciable change when *NIC* genes were compared with the less stringent (set 2, $P < 0.025$) version of DEV_{INT} (Fig. 3*B, column 2*).

Tumor expression of differentiation genes. Next, we examined the degree to which colon cancers differ from normal

epithelium in expression of transcripts with the opposite pattern (i.e., increased levels after the developmental villus transition); such genes represent independent controls for the results we have obtained thus far. To define a set of differentiation genes that appear late in mammalian intestine development, we considered all transcripts with significantly higher ($P < 0.01$ and more than eight tags) SAGE representation at E15 than at E12 and/or E13 (Fig. 4*A*). One hundred seventy-seven transcripts satisfied this requirement and constitute the set we denote as DIF_{INT}. To arrive at statistical estimates of the rate at which such late-rising genes are overexpressed in colon cancer, we formulated the following conditional probabilities:

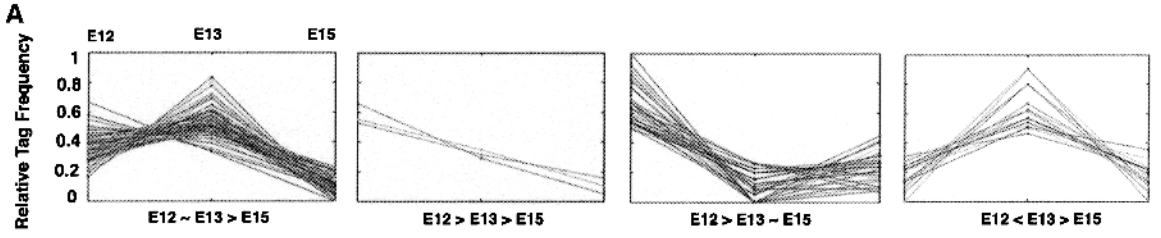
$$\Pr(x \in DIF_{INT} \mid x \in CAN_{INT}) \quad \text{(D)}$$

$$\Pr(x \in DIF_{INT} \mid x \in NC_{INT}) \quad \text{(E)}$$

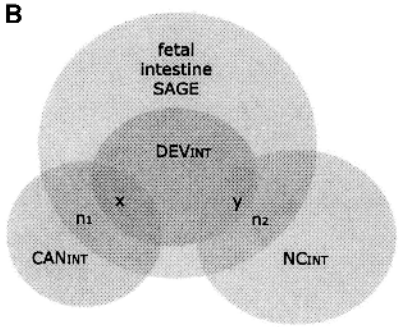
Eq. D is the conditional probability that a gene x, given that it is overexpressed in colon cancer, was activated late in fetal gut development. The second conditional is the distribution on the probability of a gene x, given that it is not overexpressed in intestinal cancer, displaying such a developmental profile.

We computed the conditional probabilities in Eqs. D and E using the same CAN_{INT} and NC_{INT} gene expression data sets described above (Table 1) and the methods applied for Eqs. A and B (Fig. 2*B*). The distribution on the likelihood of any gene increasing in expression after the fetal villous transition, given that it is overexpressed in colon cancer, lies between 0.8% and 11% (Fig. 4*B, column 1*). If a given gene is not overexpressed in colon tumors (Fig. 4*B, column 2*), its likelihood of increased expression late in gut development is between 0% and 4%. In an unpaired, nonparametric (Mann-Whitney) t test, the difference between these two conditional distributions is insignificant ($P < 0.53$). Thus, in contrast to the recapitulation of developmentally down-regulated genes in cancer, the rate of expression of differentiation genes is similar between cancerous and normal colon. However, it may be easier in cancer samples to detect developmental transcripts that are reduced or absent in normal tissue than it is to record reduced expression of differentiation markers, which could be contributed by admixed normal mucosa.

Oncofetal markers. Our analysis of recapitulation of developmental genes in cancer is limited by at least two factors: the size of currently available gene expression data sets in mouse gut development and human colon cancer and the difficulties in assigning homology between pairs of human and mouse genes. Nevertheless, the statistical basis of our analysis predicts that future studies, which might draw on more comprehensive transcriptional profiles, would yield similar trends. Meanwhile, our current results highlight some developmentally regulated genes that are reactivated in neoplasia (Table 3). Genes with these expression characteristics point to cellular functions that may be common to cancer and developing tissues. The overlap in developmental and malignant gene expression encompasses



B



C

	Test: Pr(DEVINT | CANINT) x/n1		Control: Pr(DEVINT | NCINT) y/n2	
	DEVINT set		DEVINT set	
	1 (p<0.01)	2 (p<0.25)	1 (p<0.01)	2 (p<0.25)
mArray_adCa1	7.8%	9.8%	6.7%	6.7%
mArray_adCa2	7.7%	11.5%	0.0%	6.9%
mArray_adCa3	8.7%	10.9%	2.5%	2.5%
mArray_adCa4	7.8%	11.7%	3.1%	12.5%
mArray_adCa5	8.2%	12.2%	0.0%	2.7%
mArray_GCM_colon	7.8%	9.4%	1.9%	3.8%
NC_1 vs TU_98	12.5%	14.3%	1.1%	4.5%
NC_2 vs TU_102	18.8%	22.2%	2.0%	6.1%

Figure 2. Reactivation of developmentally down-regulated genes in human colon cancer. *A,* cluster distribution of a representative fraction of 254 transcripts that show higher ($P < 0.01$) expression in the developing mouse intestine on embryonic days E12 and/or E13 relative to E15 and constitute the set denoted DEV_{INT}. These mRNAs were selected to investigate expression in human tumors. >, <, and ~ (stable) are the relationships in relative frequency of SAGE tags between developmental stages. Clusters were generated using a *k*-model-based algorithm to separate groups according to their temporal variation in expression (35). *B,* Venn diagram representation of the conditional probability distributions given by Eqs. A and B for the three categories DEV_{INT}, CAN_{INT}, and NC_{INT}, illustrating derivation of the two pertinent ratios, x/n_1 and y/n_2. *C,* table of approximations for the conditionals on the eight CAN_{INT} (two SAGE and six microarray data sets) and NC_{INT} (normal colon expression) groups. In each category, column 1 lists approximations based on strict statistical criteria ($P < 0.01$, 254 transcripts) for developmental down-regulation and column 2 lists approximations based on relaxed criteria ($P < 0.025$, 451 genes) for defining DEV_{INT}. Despite the adversity that follows from increased noise when the statistical variables are modified, the results change little, implying that the different distributions on Eqs. A and B represent true underlying biological differences.

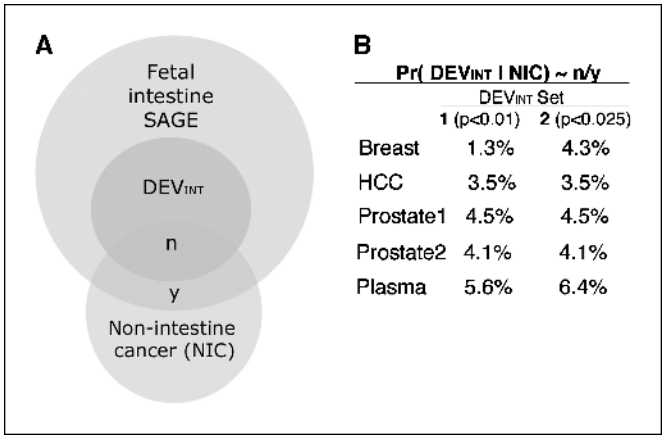


Figure 3. Tissue specificity of developmental gene reactivation in cancer. *A,* Venn diagram representation of the conditional probability distributions given by Eq. C. *B,* table of approximations for the conditional probability (Eq. C) estimated on gene expression data from five different types of human nonintestinal cancer (*NIC*). Columns 1 and 2 list, respectively, the approximations, *n/y*, based on strict ($P < 0.01$, DEV_{INT} set of 254 transcripts) and relaxed ($P < 0.025$, DEV_{INT} set of 451 transcripts) statistical variables for developmental down-regulation.

several factors associated with cellular stress response (heat shock proteins 90 and 1β, stress-induced phosphoprotein 1, and chaperonin-containing complexes), protein synthesis (ribosomal proteins L39, L3, and S26 and eukaryotic translation elongation factor 1B2), and cell cycle regulation (*CDK4* and the *Cdc47* homologue). Insulin-like growth factor II (*IGF-II*), which is known to enhance tumor growth (21), is also highly expressed in the fetal and cancerous gut. Other genes with similar expression include the paracrine growth factor midkine and an intracellular regulator of multiple Ran protein functions, *RanBP1*. These findings likely reflect the diversity of processes that mark both embryonic development and malignancy and reveal malignant reactivation of developmental genes with a range of cellular functions.

Discussion

Colorectal carcinomas are heterogeneous in their degree of differentiation and, by definition, lack the normal tissue architecture. Indeed, departure from the normal morphology is a fundamental property of cancer cells that is probably linked to invasion and other malignant behaviors. Underlying these properties is some combination of repression of terminal differentiation genes and reactivation of others associated with development of the target tissue. Previous identification of such oncofetal genes has provided both mechanistic insights in cancer biology and biomarkers that are very useful in managing a variety of human epithelial cancers. Expression profiling reveals many genes with altered expression in tumors relative to the normal tissue (11, 16). We hypothesized that a significant fraction of increased transcript levels in colon cancer reflects the developmental program of the fetal gut.

The idea that cancers share properties with developing embryos has been discussed by many authors (22, 23) and first gained currency following the embryologic studies of Waddington and Needham in the 1930s, when malignant behaviors were considered in the light of tissue organizers, morphogenetic fields, and cellular hierarchies (24, 25). Recent progress in linking epithelial tumors in general, and colon cancer in particular, to cell signaling pathways that regulate gut development and homeostasis extend these ideas significantly (2). Although many other conceptual and experimental advances have highlighted commonalities between development and cancer, the parallels have for the most part been explored at the level of single or small groups of genes and pathways. Here, we report a systematic and quantitative approach to delineate the degree of overlapping gene expression in colon cancer and development of the mammalian intestine.

We applied computational and statistical strategies to relate expression profiles from human colon cancer and the developing mouse gut. The process applied to generate the target data sets ensures that each human gene considered can be mapped with confidence to a probable murine homologue. Our results reveal that 8% to 19% of genes overexpressed in intestinal tumors had previously shown their highest expression concomitant with fetal villus morphogenesis. The frequency at which such genes are expressed in various nonintestinal tumors or the likelihood of genes not overexpressed in colon cancer having an origin in the gut developmental expression program are both much lower. For each of these trends, the results were similar when we analyzed developmental gene sets defined by different statistical criteria, which suggests that the correlations reflect the true underlying biology. Our analysis thus yields a systematic estimate of the global extent to which developmental gene expression may be recapitulated in a solid tumor and suggests that this process displays a high degree of tissue specificity.

Transcripts expressed in human cancers from various sites have been profiled much more extensively than stages in development of individual organs. This difference limits the degree to which our methods may immediately be extended to explore overlaps in gene expression between cancers and development of other tissues. However, our observations do make a testable prediction: strong correlation between activation of groups of genes in particular tumors and the prior silencing of those genes during fetal

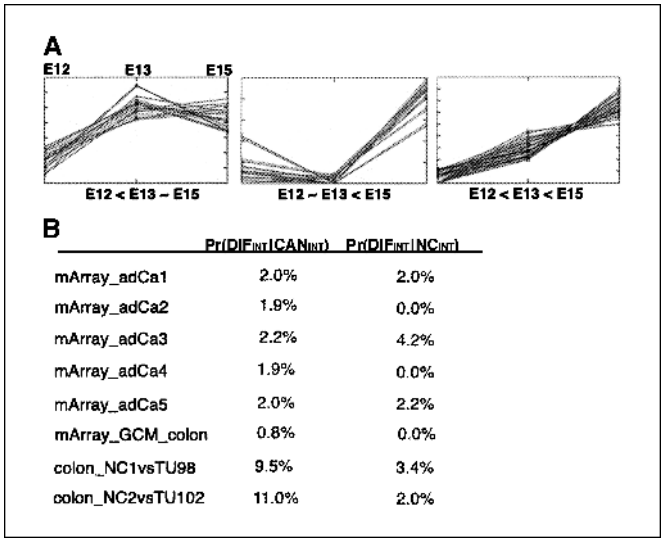


Figure 4. Expression of differentiation genes in human colon cancer. *A,* cluster representation, generated as described for Fig. 2*A*, of selected transcripts that show higher ($P < 0.01$) expression in the developing mouse gut at E15 than before; the full set of 177 genes, derived from SAGE profiling, is designated DIF_{INT}. *B,* table of approximations on the expression of differentiation genes in cancer, given by Eqs. D and E and derived from the eight CAN_{INT}, eight NC_{INT}, and DIF_{INT} gene sets.

Table 3. Transcripts overexpressed in human colon cancer, as detected by microarray or SAGE analysis, and also down-regulated during mouse intestine organogenesis

LocusLink (human)	Unigene (mouse)	Expression pattern in fetal gut	Gene description
Cancer gene overexpression detected by microarray			
641	12932	E12 ~ E13 > E15	Bloom syndrome (*BLM*)
22948	1813	E12 > E13 > E15	Chaperonin containing TCP1 s5 epsilon (*CCT5*)
3320	1843	E12 > E13 ~ E15	Heat shock 90-kDa protein 1, α
3192	2115	E12 ~ E13 > E15	Breakpoint cluster region
65108	2769	E12 < E13 > E15	MARCKS-like protein
23204	29924	E12 ~ E13 > E15	ADP-ribosylation factor-l6 interacting protein
7329	3268	E12 < E13 > E15	Ubiquitin-conjugating enzyme E2I
10963	4540	E12 < E13 > E15	Stress-induced-phosphoprotein 1
4869	6343	E12 > E13 > E15	Nucleolar phosphoprotein B23
1019	6839	E12 < E13 > E15	Cyclin-dependent kinase 4 (*CDK4*) gene
3418	246432	E12 ~ E13 > E15	Isocitrate dehydrogenase 2 (*NADP+*), mitochondrial
3608	21534	E12 > E13 ~ E15	Nuclear factor NF45 mRNA (*ILF2*)
4176	18923	E12 ~ E13 > E15	DNA replication factor CDC47 homologue (*MCM7*)
4192	906	E12 < E13 > E15	Midkine (neurite growth-promoting factor 2)
5902	3752	E12 ~ E13 > E15	RAN binding protein 1 (*RANBP1*)
6428	6787	E12 > E13 ~ E15	Pre-mRNA splicing factor SRP20
6749	219793	E12 < E13 > E15	SSRP1 High mobility group box
7001	42948	E12 < E13 > E15	Thiol-specific antioxidant protein
7045	14455	E12 > E13 ~ E15	Transforming growth factor β induced gene (*BIGH3*)
Cancer gene overexpression detected by SAGE			
3326	2180	E12 > E13 > E15	Heat shock protein 1, β
1933	2718	E12 > E13 > E15	Eukaryotic translation elongation factor 1B2
6170	30478	E12 > E13 > E15	Ribosomal protein L39
6122	3486	E12 < E13 > E15	Ribosomal protein L3
6231	372	E12 > E13 ~ E15	Ribosomal protein S26
3481	3862	E12 > E13 > E15	*IGF-II*

NOTE: Fetal expression patterns, archived at http://genome.dfci.harvard.edu/GutSAGE, were recorded on gestational days E12, E13, and E15. ~, roughly equal expression.

patterning of the same tissue. Notably, our definitions do not lead to the claim that there are more cancer transcripts than noncancer transcripts of developmental origin; many genes reduced late in development are expressed to varying degrees in the normal adult tissue. Nor do we imply that most transcripts overexpressed in cancer belong to the developmental program. There are undoubtedly many modes of aberrant gene activation in tumors, of which developmental gene reactivation is only one.

Tissue differentiation during development is largely under epigenetic control. It follows that there may be two classes of epigenetic modification pertinent to this discussion. One category is represented in genes that are never accessible to the transcription machinery in a particular cell type and accordingly never expressed therein. The second type of modification occurs in genes after their transient embryonic expression and such changes may be especially amenable to reversal in malignancy. Considerable experimental evidence indicates that epigenetic alteration, including DNA methylation, is important in tumor initiation and progression (26, 27); investigation has traditionally emphasized inactivation of tumor suppressor genes, usually by promoter hypermethylation (28). In contrast, reactivation of developmentally regulated genes may result from the aberrant hypomethylation observed in many tumor types, including colon cancer (29), and likely contributes to malignant behaviors.

Historically, oncofetal proteins have been identified individually. Our computational analysis of malignant and developmental expression profiles represents a strategy to identify new candidates and reveals tumor reactivation of fetal genes with a wide range of cellular functions. A few gene classes are prominent, including factors associated with cellular stress response, cell cycle regulation, and protein synthesis. Although increased expression of ribosomal protein genes is recognized as a feature of human cancers (30, 31), the underlying significance is uncertain. Conversely, monoallelic loss-of-function mutations in gene loci for ribosomal proteins were reported to a high and unexpected degree in tumors in zebrafish (32). These observations support the idea that dysregulated ribosome biogenesis, whether by gain or loss of gene functions, may be tumorigenic (33). We find that developmental transitions are accompanied by significant modulation in expression of selected genes involved in assembly of the 40S and 60S ribosomal subunits (http://genome.dfci.harvard.edu/GutSAGE). Thus, certain aspects of the ribosome or, alternatively, nonribosomal functions of the genes in question, may be common to developmental and malignant cellular processes. *IGF-II*, which is known to enhance tumor growth and suppress apoptosis (21), is also highly expressed in the fetal and cancerous gut. Loss of genomic imprinting, with resulting abnormal activation of the normally silent maternal *Igf2* allele in all cells, is strongly associated with a

risk of developing colon cancer (34). Besides their value in cancer diagnosis and surveillance, oncofetal factors could also be good therapeutic targets. If they are expressed on the tumor cell surface or make a material contribution to the malignant phenotype, then specific antibodies or drugs may be developed in anticipation of limited toxicity as the target is absent from the normal adult tissue.

Acknowledgments

Received 3/1/2005; revised 7/7/2005; accepted 7/22/2005.

Grant support: Robert Black Charitable Foundation and NIH grant R01DK61139.

The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked *advertisement* in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

We thank Maina Lepourcelet and Li Cai for their contributions toward early phases of this work and members of the Shivdasani laboratory for helpful discussions.

References

1. Lum L, Beachy PA. The Hedgehog response network: sensors, switches, and routers. Science 2004;304:1755–9.
2. Sancho E, Batlle E, Clevers H. Signaling pathways in intestinal development and cancer. Annu Rev Cell Dev Biol 2004;20:695–723.
3. Vogelstein B, Kinzler KW. Cancer genes and the pathways they control. Nat Med 2004;10:789–99.
4. Garrett PE, Kurtz SR. Clinical utility of oncofetal proteins and hormones as tumor markers. Med Clin North Am 1986;70:1295–306.
5. Duffy MJ. Evidence for the clinical use of tumour markers. Ann Clin Biochem 2004;41:370–7.
6. Kho AT, Zhao Q, Cai Z, et al. Conserved mechanisms across development and tumorigenesis revealed by a mouse development perspective of human cancers. Genes Dev 2004;18:629–40.
7. Stappenbeck TS, Wong MH, Saam JR, Mysorekar IU, Gordon JI. Notes from some crypt watchers: regulation of renewal in the mouse intestinal epithelium. Curr Opin Cell Biol 1998;10:702–9.
8. Potten CS, Booth C, Pritchard DM. The intestinal epithelial stem cell: the mucosal governor. Int J Exp Pathol 1997;78:219–43.
9. Velculescu VE, Zhang L, Vogelstein B, Kinzler KW. Serial analysis of gene expression. Science 1995;270: 484–7.
10. Lepourcelet M, Tou L, Cai L, et al. Insights into developmental mechanisms and cancers in the mammalian intestine derived from serial analysis of gene expression and study of the hepatoma-derived growth factor (HDGF). Development 2005;132:415–27.
11. Rhodes DR, Yu J, Shanker K, et al. ONCOMINE: a cancer microarray database and integrated data-mining platform. Neoplasia 2004;6:1–6.
12. Zhang L, Zhou W, Velculescu VE, et al. Gene expression profiles in normal and cancer cells. Science 1997;276:1268–72.
13. Alon U, Barkai N, Notterman DA, et al. Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. Proc Natl Acad Sci U S A 1999; 96:6745–50.
14. Notterman DA, Alon U, Sierk AJ, Levine AJ. Transcriptional gene expression profiles of colorectal adenoma, adenocarcinoma, and normal tissue examined by oligonucleotide arrays. Cancer Res 2001;61:3124–30.
15. Ramaswamy S, Tamayo P, Rifkin R, et al. Multiclass cancer diagnosis using tumor gene expression signatures. Proc Natl Acad Sci U S A 2001;98:15149–54.
16. Lal A, Lash AE, Altschul SF, et al. A public database for gene expression in human cancers. Cancer Res 1999; 59:5403–7.
17. Ma XJ, Salunga R, Tuggle JT, et al. Gene expression profiles of human breast cancer progression. Proc Natl Acad Sci U S A 2003;100:5974–9.
18. Chen X, Cheung ST, So S, et al. Gene expression patterns in human liver cancers. Mol Biol Cell 2002;13: 1929–39.
19. Singh D, Febbo PG, Ross K, et al. Gene expression correlates of clinical prostate cancer behavior. Cancer Cell 2002;1:203–9.
20. De Vos J, Thykjaer T, Tarte K, et al. Comparison of gene expression profiling between malignant and normal plasma cells with oligonucleotide arrays. Oncogene 2002;21:6848–57.
21. Furstenberger G, Senn HJ. Insulin-like growth factors and cancer. Lancet Oncol 2002;3:298–302.
22. Uriel J. Fetal characteristics of cancer. 3rd ed. In: Becker FF, editor. Cancer, a comprehensive treatise. New York: Plenum Press; 1975.
23. Rubin H. Cancer as a dynamic developmental disorder. Cancer Res 1985;45:2935–42.
24. Needham J. New advances in the chemistry and biology of organized growth. Proc R Soc Lond B Biol Sci 1936;29:1577–626.
25. Waddington CH. Cancer and the theory of organizers. Nature 1935;135:606–8.
26. Gaudet F, Hodgson JG, Eden A, et al. Induction of tumors in mice by genomic hypomethylation. Science 2003;300:489–92.
27. Mihich E, Jaenisch R. Sixteenth Annual Pezcoller Symposium: stem cells and epigenesis in cancer. Cancer Res 2004;64:8474–7.
28. Jones PA, Baylin SB. The fundamental role of epigenetic events in cancer. Nat Rev Genet 2002;3: 415–28.
29. Goelz SE, Vogelstein B, Hamilton SR, Feinberg AP. Hypomethylation of DNA from benign and malignant human colon neoplasms. Science 1985;228:187–90.
30. Loging WT, Reisman D. Elevated expression of ribosomal protein genes L37, RPP-1, and S2 in the presence of mutant p53. Cancer Epidemiol Biomarkers Prev 1999;8:1011–6.
31. Kondoh N, Shuda M, Tanaka K, Wakatsuki T, Hada A, Yamamoto M. Enhanced expression of S8, L12, L23a, L27 and L30 ribosomal protein mRNAs in human hepatocellular carcinoma. Anticancer Res 2001;21: 2429–33.
32. Amsterdam A, Sadler KC, Lai K, et al. Many ribosomal protein genes are cancer genes in zebrafish. PLoS Biol 2004;2:690–8.
33. Ruggero D, Pandolfi PP. Does the ribosome translate cancer? Nat Rev Cancer 2003;3:179–92.
34. Cui H, Cruz-Correa M, Giardiello FM, et al. Loss of IGF2 imprinting: a potential marker of colorectal cancer risk. Science 2003;299:1753–5.
35. Cai L, Huang H, Blackshaw S, Liu JS, Cepko C, Wong WH. Clustering analysis of SAGE data using a Poisson approach. Genome Biol 2004;5:R51.