GASTROENTEROLOGY 2002;122:1467–1482

Novel Genes and Functional Relationships in the Adult Mouse Gastrointestinal Tract Identified by Microarray Analysis

MICHAEL D. BATES,* CHRISTOPHER R. ERWIN,† L. PHILIP SANFORD,* DAN WIGINTON,§ JORGE A. BEZERRA,* LYNN C. SCHATZMAN,* ANIL G. JEGGA,‖ CATHY LEY-EBERT,§ SARAH S. WILLIAMS,‖ KRIS A. STEINBRECHER,*,§ BRAD W. WARNER,† MITCHELL B. COHEN,* and BRUCE J. ARONOW§,‖

*Division of Gastroenterology, Hepatology and Nutrition, †Division of Pediatric Surgery, §Division of Developmental Biology, ‖Division of Pediatric Informatics, Children's Hospital Medical Center, University of Cincinnati College of Medicine, Cincinnati, Ohio

Background & Aims: **A genome-level understanding of the molecular basis of segmental gene expression along the anterior-posterior (A-P) axis of the mammalian gastrointestinal (GI) tract is lacking. We hypothesized that functional patterning along the A-P axis of the GI tract could be defined at the molecular level by analyzing expression profiles of large numbers of genes.** ***Methods:*** **Incyte GEM1 microarrays containing 8638 complementary DNAs (cDNAs) were used to define expression profiles in adult mouse stomach, duodenum, jejunum, ileum, cecum, proximal colon, and distal colon. Highly expressed cDNAs were classified based on segmental expression patterns and protein function.** ***Results:*** **571 cDNAs were expressed 2-fold higher than reference in at least 1 GI tissue. Most of these genes displayed sharp segmental expression boundaries, the majority of which were at anatomically defined locations. Boundaries were particularly striking for genes encoding proteins that function in intermediary metabolism, transport, and cell-cell communication. Genes with distinctive expression profiles were compared with mouse and human genomic sequence for promoter analysis and gene discovery.** ***Conclusions:*** **The anatomically defined organs of the GI tract (stomach, small intestine, colon) can be distinguished based on a genome-level analysis of gene expression profiles. However, distinctions between various regions of the small intestine and colon are much less striking. We have identified novel genes not previously known to be expressed in the adult GI tract. Identification of genes coordinately regulated along the A-P axis provides a basis for new insights and gene discovery relevant to GI development, differentiation, function, and disease.**

The mammalian digestive system develops with respect to several axes that guide patterning: anterior-posterior (A-P; cranial-caudal, proximal-distal), left-right, dorsal-ventral, and mucosal-serosal (radial).[1,2] Later in development, the crypt-villus (vertical) axis is formed as a basis for epithelial function and renewal. The A-P axis is particularly important because it is necessary for the regional specificity of gastrointestinal (GI) function. Although anatomic and physiologic definitions and boundaries of the various portions of the GI tract are well established, the molecular mechanisms underlying the development and maintenance of regional specification are largely unknown.

Over the past few years, techniques allowing the simultaneous measurement of the expression of hundreds or thousands of genes have been developed.[3] These techniques use arrays of genes on a medium, such as filters or glass slides. The most recent iteration of this approach uses DNA microarrays containing thousands of gene sequences spotted or synthesized on glass slides (or chips), to which labeled complementary DNA (cDNA) or complementary RNA samples are hybridized. At present, as many as 8000–12,000 genes can be represented on 1 chip. Given current estimates of 35,000–42,000 genes in a mammalian genome, this means that the level of expression of 20%–25% of mammalian genes can be measured in 1 experiment.

We have used the power of microarray-based genomics to examine patterns of gene expression in the adult mouse GI tract. Specifically, we were interested in understanding the molecular basis of regional specification of function along the A-P axis of the adult GI tract. We hypothesized that the molecular definition of this axis could be confirmed by analyzing the expression and function of large numbers of genes. This approach re-

Abbreviations used in this paper: Amn, amnionless gene; A-P, anterior-posterior; EST, expressed sequence tag; GI, gastrointestinal; poly A^+ RNA, polyadenylated RNA; P1, postnatal day 1.

0016-5085/02/$35.00
doi:10.1053/gast.2002.32975

vealed novel insights into the characteristics of gene expression along the A-P axis of the adult GI tract, including striking gene expression and functional transitions between the stomach and small intestine, and between the small intestine and colon, that parallel anatomically and physiologically defined transitions. The overall patterns observed support a model of transcriptional regulation of boundaries of gene expression in the GI tract. Finally, we have also shown the use of this approach for promoter analysis of coordinately regulated genes and for gene discovery relevant to GI function.

Materials and Methods

Animals, Tissue Samples, and RNA Preparation

Adult GI tissues were obtained from 6- to 8-week-old male C57BL/6 mice (Jackson Laboratories, Bar Harbor, ME) as part of an institutional consortium effort to develop a generalized mouse gene expression database. This database includes a wide range of developing, normal adult, and diseased adult tissues, including brain, dorsal root ganglion, heart, lung, kidney, liver, reproductive and endocrine organs, immune tissues, muscle, and skin. Seven adult GI tissues were selected for inclusion in the database. The stomach was represented by the hindstomach, which is anatomically distinct in the mouse from the forestomach. Duodenum was defined as the section of small intestine between the pylorus and the ligament of Treitz. Jejunum and ileum were defined as the proximal and distal thirds, respectively, of the small intestine between the ligament of Treitz and cecum. Proximal colon and distal colon were defined as the proximal 40% and distal 60%, respectively, of the colon, exclusive of the cecum. These tissues include the variety of cell lineages (epithelial, mesenchymal, immune, neuronal) present in the wall of each tissue, but mesenteries were removed. For each sample, 3–6 mice were killed by carbon dioxide inhalation according to institutional and national animal care guidelines, and tissues were rapidly dissected and snap frozen in liquid nitrogen. Frozen tissues were ground to powder by using a ribonuclease-free mortar and pestle, and total RNA was prepared by using TriZOL (Gibco/Life Technologies, Rockville, MD), following the manufacturer's protocol. Total RNA was reprecipitated with ethanol/sodium acetate and resuspended in diethylpyrocarbonate-treated water. Polyadenylated (poly A^+) RNA was prepared from total RNA by using Oligotex (Qiagen, Valencia, CA). Poly A^+ RNA was then quantitated by using RiboGreen dye (Molecular Probes, Eugene, OR) and checked for RNase degradation by agarose gel electrophoresis and the use of an Agilent Bioanalyzer 2100 (Agilent Technologies, Palo Alto, CA). A total of 600 ng of poly A^+ RNA (50 ng/μL) for each tissue was submitted for cDNA labeling and microarray hybridization.

Fluorescent Probe Preparation and Microarray Hybridization

Probe preparation and microarray hybridization were performed by Incyte Genomics (Palo Alto, CA). Labeled cDNA was prepared from the poly A^+ RNA sample from each GI tissue by using nucleotides labeled with the fluorescent dye Cy5. Labeled cDNAs were then subjected to competitive hybridization to mouse GEM1 microarrays, which contained spotted cDNAs corresponding to 8638 sequence-verified expressed sequence tag (EST) clones (500–5000 nucleotides) from the Integrated Molecular Analysis of Genomes and their Expression I.M.A.G.E. Consortium with a low redundancy rate with respect to their corresponding gene products (http://gastro.chmcc.org). The sequences used for highly homologous genes on the arrays were examined because the ability to distinguish between family members is dependent on the specific sequences represented and the hybridization stringency used. For all samples, hybridizations were performed in competition with Cy3-labeled cDNA from whole postnatal day 1 (P1) mouse. This reference sample was chosen based on independent tests of its ability to generate reproducible competitive hybridizations from 3 different P1 mouse isolates (data not shown). Additional control experiments using highly divergent samples indicated little or no differences in hybridization intensity if the fluorescent dyes were reversed (data not shown). The use of a single common reference messenger RNA allows for expression profile comparisons among a variety of developing and adult tissues. Fluorescence intensity analyses and background subtraction were performed by using an Axon Instruments scanner and their GenePix software (Union City, CA).

Although cDNA microarrays generally offer outstanding robustness, signal intensity, and sensitivity, an important question is the extent to which highly homologous gene sequences can be distinguished given the target sizes and hybridization stringency used. Both extent of homology of the sequences represented on the arrays and differences in relative abundance in the sample may mask the ability to resolve closely related gene products. As homology increases, the need for additional sequence-specific validation methods becomes imperative. However, we have not found any clear instances of 2 different genes represented on the arrays by sequences of high homology that also display similar expression profiles.

Microarray Data Analyses

Primary quantitative data, spot geometry, and background fluorescence were examined by using Incyte Gemtools software. Defective cDNA spots (i.e., those with irregular geometry, scratched, or containing $<40\%$ of the area compared with average) were eliminated from the dataset. All datasets were normalized by using a balancing coefficient of the median of all Cy5 channel measurements divided by the median of all Cy3 channel measurements. Each microarray

contained 192 control genes present as either nonmammalian single gene "spikes" or complex targets consisting of pools of genes expressed in most cell types, and each experimental messenger RNA sample was augmented with incremental amounts of nonmammalian gene RNA (2-fold, 4-fold, 16-fold, etc.) to permit assessment of the dynamic range attained within each microarray. Less than 2-fold variation was observed across the microarray series with respect to the 192 control genes (data not shown), providing additional support for the feasibility of interarray comparisons to detect genes regulated in a tissue-specific manner. Such reproducibility was also observed for genes highly expressed in each tissue. In 6 of the 7 tissues studied there was 94% or greater correspondence between the cDNAs found to be highly expressed in the 2 hybridizations; reproducibility was 87% for stomach. Second-stage data analyses were performed by using GeneSpring software (Silicon Genetics, Redwood City, CA). A total of 571 cDNAs corresponding to genes highly expressed in the adult GI tract were identified based on expression that was 2-fold or higher compared with whole P1 mouse in both hybridizations for at least 1 adult GI tissue.

These 571 highly expressed cDNAs were clustered according to their expression profiles along the GI A-P axis by using hierarchical tree clustering[4] and K-means[5] algorithms as implemented in the GeneSpring program. Hierarchical clustering allows unbiased grouping of genes based on relative expression across the sample series. Clustering results will differ depending on whether they are performed using data based on $\log_2$-transformed ratios of expression between the tissue of interest and reference, or using untransformed ratios. K-means analysis places gene expression profiles into a predefined number of clusters based on relative expression across the sample series. We performed clustering of the dataset by using several different normalization methods. The use of raw ratios of hybridization vs. reference or the $\log_2$ of these ratios provides an assessment of genes based on their levels of expression and allows the identification of genes that are highly expressed in the GI tract but do not display segmental specificity. For other analyses, normalization of raw ratios or the $\log_2$ of these ratios to their median expression across the tissue series ("each gene normalized in GeneSpring") was performed. This approach is particularly useful for assessing segment-specific expression patterns irrespective of the absolute expression level.

To allow application of biologically relevant constraints to the analysis of expression profiles, we also developed an alternative method to group expression profiles based on the tissues where they are highly expressed relative to reference. This analysis was performed by using $\log_2$-transformed data exported from GeneSpring into tabular format for grouping of genes with similar expression patterns by using Microsoft Excel (Redmond, WA). Thus, the list of 571 highly expressed cDNAs was sorted based on the tissue(s) in which they were highly expressed, and secondarily based on the differences in expression relative to reference (on a $\log_2$ scale).

Identification and Functional Classification of Highly Expressed Genes

cDNAs to which there was increased hybridization in GI tissues as compared with reference were identified and classified by encoded protein function if possible. The Incyte GEM1 microarray contains cDNA clones corresponding to identified genes, cDNAs from ESTs, and other unannotated sequences. GenBank accession numbers for annotated genes and ESTs were used to confirm or identify the gene and encoded protein by using the UniGene assembly database for Mus musculus (http://www.ncbi.nlm.nih.gov/UniGene/Mm.Home.html) and Basic Local Alignment Search Tool (Nucleotide-Nucleotide) (BLASTN) searches (http://www.ncbi.nlm.nih.gov/BLAST/) vs. the nonredundant nucleotide database over a period of March–August 2001. We found very few clones whose annotations required revision. For unannotated EST sequences, UniGene and BLAST searches were performed from within GeneSpring. By using these methods, we were able to identify the gene and encoded protein for approximately one third of the ESTs and unannotated sequences. In addition, we were able to identify a known or predicted gene for 14 of 20 unknown ESTs by using the mouse genome assembly within the Celera Discovery System (Celera Genomics, Rockville, MD) (described later). We classified the genes and encoded proteins by biochemical function. To date, there is no standard classification scheme for protein function. Based on a variety of reported classification schemes, including those

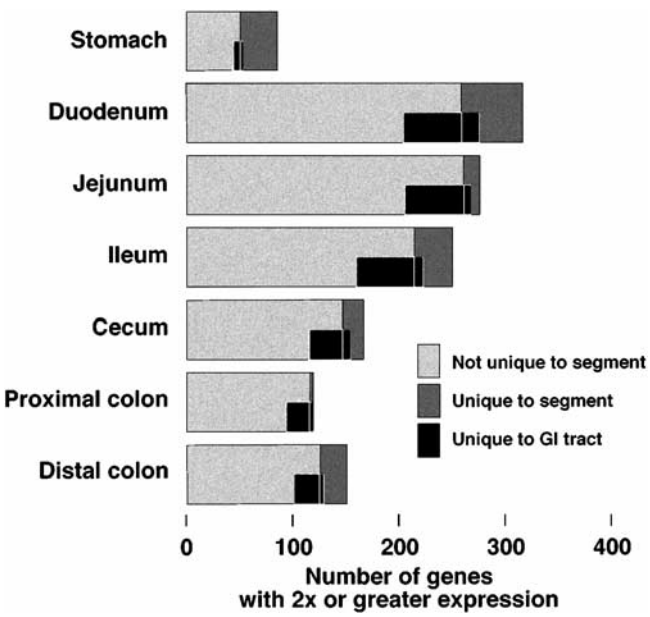

Figure 1. Numbers of genes highly expressed in specific segments of the adult mouse GI tract. Graph shows the numbers of genes highly expressed in each segment of the GI tract, including the number of genes whose increased expression is limited to that particular segment (unique to segment). □ , Numbers of genes with high expression that is not unique to individual tissue segment; ■, Numbers with high expression uniquely in the individual tissue segment; ■. Numbers with high expression uniquely in GI tissues.

Functional Category	Identification	Accession numbers: Incyte	Accession numbers: GenBank	Only in GI	Expression pattern
Cell-Cell Communication	Apolipoprotein B-100	676410	AA209065		
	Apolipoprotein B-100	723038	AA254389		
	Fibroblast growth factor 15	479758	AA051675	X	
	Guanylin	889440	AA498457	X	
	Indian hedgehog	699460	AA245525	X	
	Somatostatin	337219	W20888		
	Trefoil factor 1	404550	W83072	X	
	Trefoil factor 2	314078	W09829		
	Trefoil factor 3	762404	AA273366	X	
DNA/RNA Processes	Cdx1	633635	AA184566	X	
	Cdx2	697874	AA238566	X	
	Krüppel-like factor 5 (Intestinal Krüppel-like factor)	444844	AA016731	X	
	Max protein	793537	AA415602	X	
	Max protein	481641	AA060226	X	
Immunity/Defense	β2 microglobulin	572542	AA109951		
	H-2, class I, L region	694651	AA221044		
	H-2, class I, L region	920425	AA538511		
	H-2, class II antigen A, alpha	747378	AA272807		
	H-2, class II antigen A, beta 1	618271	AA175329		
	IgA heavy chain	596604	AA146478		
	IgA heavy chain	749660	AA388939		
	IgA heavy chain	620655	AA177218		
Intermediary Metabolism	Adenosine deaminase	387459	W65788.1	X	
	Carbonic anhydrase 1	922088	AA512372		
	Carbonic anhydrase 2	579391	AA122925		
	Colipase	481341	AI385475	X	
	Intestinal alkaline phosphatase	493555	AA087125	X	
	Monoamine oxidase B	680958	AA241899		
Misc. Cell Processes	Caspase 1	550766	AI326615	X	
	Intestinal fatty acid binding protein 2	679661	AA245078	X	
	Proprotein convertase subtilisin/kexin type 3	583759	AI326760	X	
	Protein disulfide isomerase	354215	W44037.1		
	Protein disulfide isomerase	722782	AA254546	X	
	Protein kinase C zeta	442856	AI894049	X	
Structural/Cytoskeletal	Claudin-2	888602	AA498630		
	Claudin-3	441695	AA013960		
	Claudin-7	717135	AA266234		
	Claudin-15	466280	AA033362.1	X	
	Galectin-6	717208	AA404174	X	
	Galectin-9	680815	AA250039		
Transport	Apolipoprotein A-IV	492937	AA097421	X	
	Apolipoprotein C-II	736992	AA271959		
	Apolipoprotein C-III	356196	W50759		
	Polymeric Ig receptor	737364	AA277571		
Unknown/Other	Crp-ductin	888707	AA498773		
	Reg IIIb/Pancreatitis-associated protein 1	873502	AA473899	X	

Figure 2.

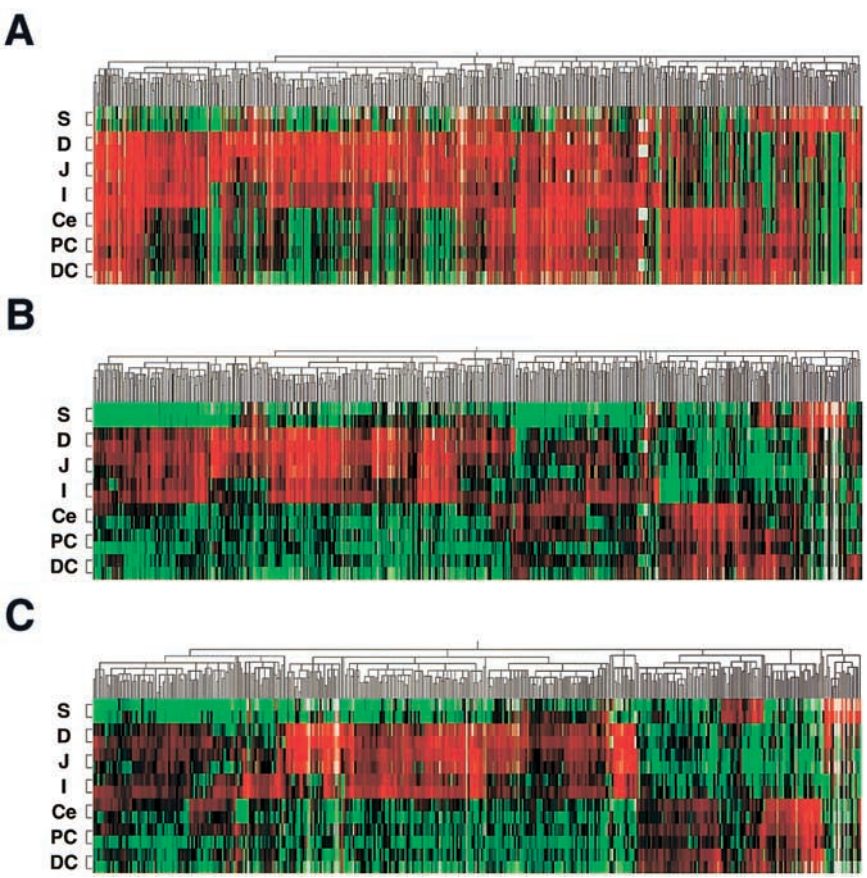

Figure 3. Hierarchical clustering of genes highly expressed in adult mouse GI tissues. A total of 571 cDNAs with expression 2-fold or greater relative to reference in replicate hybridizations for 1 or more of the 7 adult GI tissues were identified. Expression values were subjected to hierarchical clustering, as described in the Materials and Methods section, by using a minimum distance value of 0.001 and a separation ratio of 0.5. Data for replicate hybridizations are shown. Colors are graded with red indicating increased expression and green indicating low expression relative to reference. White tiles indicate microarray elements that did not pass quality control as described in the Materials and Methods section. (*A*) Hierarchical clustering applied to $\log_2$-transformed ratios relative to reference, using Pearson correlation. (*B*) Hierarchical clustering using $\log_2$-transformed ratios for each gene normalized to its median level of expression among the GI tissues ("each gene normalized"). The resulting hierarchy is identical to that in panel *A*, but the colors more clearly depict expression differences between GI tissues. (*C*) Hierarchical clustering with standard correlation applied to relative expression ratio values for each gene normalized to its median level of expression among the GI tissues ("each gene normalized"; http://www.sigenetics.com/cgi/SiG.cgi/Products/GeneSpring/GSFAQ.smf). S, stomach; D, duodenum; J, jejunum; I, ileum; Ce, cecum; PC, proximal colon; DC, distal colon.

Figure 2. Expression of selected genes in the adult mouse gastrointestinal tract. Clones are grouped by functional category of the corresponding protein and identified by protein identity and accession numbers. Clones that are not expressed greater than 2-fold above reference in any normal adult non-GI tissue in our mouse gene expression database are indicated. Expression profiles are shown by colored squares, with 2 squares (1 for each microarray hybridization) for each tissue. Colors are graded with red indicating increased expression and green indicating low expression relative to reference. S, stomach; D, duodenum; J, jejunum; I, ileum; Ce, cecum; PC, proximal colon; DC, distal colon.

for the results of the human genome project,[6,7] we classified the encoded proteins into the following categories: cell-cell communication, DNA/RNA processes (including transcription factors), immunity/defense, metabolism, miscellaneous cell processes (e.g., protein processing), structure/cytoskeleton, transport, and unknown/other (including multifunctional proteins and proteins of unknown function).

A small group of unknown and unidentified clones with interesting expression profiles was selected for further analysis by using the Celera Discovery System and Celera's associated databases (Celera Genomics). In particular, we made use of mouse genome sequence data compiled by Celera. Gene prediction by Celera was performed by using the Otto method as described.[7]

Consensus Gene Regulatory Sequences

To identify consensus gene expression regulatory sequences that are present in coordinately regulated genes, we generated lists of cDNAs giving highly similar expression profiles across our mouse gene expression database. To begin to address segmental regulation of gene expression along the A-P axis of the GI tract, we focused on genes with boundaries of expression between adjacent GI segments. Of the 8638 cDNAs represented on the mouse GEM1 microarray, we found 175 whose expression was restricted to the GI tract (increased 2-fold above the median level of "each gene normalized" expression in at least 1 adult GI tissue but no non-GI tissue). We selected cDNAs with similar expression profiles (i.e., maximal expression in the small intestine with peak expression in the ileum [SI group], or maximal expression in the colon with peak expression in the cecum [LI group]). Selection was confirmed by K-means clustering analysis of the 175 GI-restricted profiles. We found 3 genes in the SI group and 5 in the LI group. The LI group was further refined by removal of 2 clones for which full-length mouse gene sequence was not available in the Celera mouse genome assembly. Regions of phylogenetic conservation between mouse and human orthologues were identified for 4 of these genes by alignment of Celera-identified gene sequences using the program PipMaker (http://bio.cse.psu.edu/pipmaker/).[8] Consensus *cis*-regulatory/transcription factor binding elements in the proximal 5′ 2-kilobase upstream regions of each gene were identified by using the program MatInspector Professional 5.0 (Genomatix Software GmbH, Munich, Germany; http://genomatix.gsf.de/), which makes use of the TRANSFAC database.[9]

Reverse Transcription–Polymerase Chain Reaction Analysis

For analysis of *amnionless* gene expression, cDNA was prepared from poly A^+ RNA from GI tissues by using reverse transcriptase (SuperScript II; Gibco BRL), with parallel preparations lacking reverse transcriptase as negative controls. Oligonucleotide primers designed using the program MacVector (Oxford Molecular Group, Oxford, England) to amplify a 306-base pair (bp) segment of the *amnionless* coding region were: forward, 5′-AGAAGGTGGACATCTTGGACATTG-3′; reverse, 5′-ATGGTAACAGCACTTGCGGC-3′. Polymerase chain reaction was performed in 50-μL reactions containing 2.5 units of *Taq* DNA polymerase (Qiagen), 200 μmol/L of each deoxyribonucleotide, 20 pmol of each primer, and 2 mmol/L Mg^{2+}. Melting, annealing, and extension times were each 1 minute, for a total of 30 cycles. The annealing temperature was 51°C. Polymerase chain reaction products were resolved on 1.8% agarose gels by electrophoresis. Presence of mouse glyceraldehyde-3-phosphate dehydrogenase sequences was detected by using oligonucleotide primers obtained from Clontech (Palo Alto, CA).

Data Archive

Gene identities, expression data, cluster groups, and the functional categories of the dynamically regulated genes are available on our microarray database web server (http://gastro.chmcc.org).

Results

Identification of Genes Highly Expressed in Adult Gastrointestinal Tissues

To define genes important for GI function, we identified arrayed cDNAs that hybridized to adult mouse GI cDNA at a level 2-fold or greater than reference. We found 571 cDNAs and ESTs with increased expression of their corresponding genes in replicate hybridizations for 1 or more of the 7 adult GI tissues as compared with reference. Among the GI tissues, the largest number of highly expressed genes was found in the duodenum (n = 332); this number decreases moving distally in the small intestine (Figure 1). The smallest number of highly expressed genes was found in the stomach (n = 86), though it had the largest percentage of genes whose high expression was unique to this tissue within the GI tract (43%). By contrast, the proximal colon had the next lowest number of highly expressed genes (n = 119), only 2 of which were unique in their high expression in this tissue.

The 571 cDNAs and ESTs with high expression in the GI tract are generally not highly expressed in non-GI tissues as compared with the other 8067 genes repre-

sented on the mouse GEM1 microarray (data not shown). Of this set, 114 cDNAs/ESTs (20%) were highly expressed only in GI tissues and not the other tissues represented in the database (defined as less than 2-fold compared with whole P1 mouse in any hybridizations for non-GI tissues). Of these, 54 (47%) encode unknown genes. The number of genes highly expressed in the various segments of the GI tract but not highly expressed in non-GI tissues, is indicated in Figure 1. Examples of genes not highly expressed in non-GI tissues are listed in Figure 2.

The highly expressed GI genes encode proteins that participate in a wide variety of normal and pathophysiologic processes in the GI tract (Figure 2) and that display a diverse array of expression profiles along the A-P axis of the GI tract. Such processes include digestion and nutrient processing, mucosal structure and integrity, and enzymes required for biogenic amine metabolism. Processes related to pathophysiology include mucosal immunity (immunoglobulins, major histocompatibility proteins), mucosal restitution after injury (the 3 trefoil factors and a putative trefoil receptor, Crpductin[10]), and carcinogenesis (Cdx2, Max protein). Note that several genes are represented by more than 1 cDNA, but similar expression profiles are observed for each. In addition, there are a number of unknown genes represented, providing a basis for efforts toward gene discovery (see later).

Furthermore, the microarrays were able to distinguish the expression of members of various gene/protein families, such as the 3 trefoil factors, because of the distinct cDNA sequences represented on these arrays. Cross-comparisons of the trefoil factor cDNAs used (full-length I.M.A.G.E. Consortium EST clones in each case) reveal $<50\%$ identity at the nucleotide level (data not shown). The expression profiles characterized for the 3 trefoil factors were distinct: trefoil factor 1 and trefoil factor 2 in more proximal segments and trefoil factor 3 in more distal segments. In addition, the expression of trefoil factor 1 and trefoil factor 3 is restricted to the GI tract, unlike that of trefoil factor 2 (Figure 2). Thus, these related family members can be distinguished by these microarrays.

Dynamic Segmental Regulation of Gene Expression in the Adult Mouse Gastrointestinal Tract

We used 3 different approaches to examine the patterns of expression of these 571 genes within the GI tract. First, we used hierarchical clustering[4] to group the expression profiles (Figure 3). Expression in each tissue is indicated by a color on a red-black-green scale, with red indicating higher expression and green indicating lower expression. As expected, clustering results differed when we used data based on $\log_2$-transformed ratios of expression between the tissue of interest and reference (Figures 3*A–B*), vs. untransformed ratios (Figure 3*C*). No differences in the final gene clustering hierarchy were observed if $\log_2$-transformed ratio data were normalized to the median level of expression across the GI tissues (compare Figures 3*A* and 4*B*). These 3 trees reveal different aspects of the dataset. For example, expression profiles relative to reference are useful for identifying genes highly expressed across the GI tract (Figure 3*A*), whereas the "each gene" normalization provides a more robust display of changes in the expression levels of genes with highly regulated expression within the GI tract (Figures 3*B–C*). Overall, hierarchical clustering analysis reveals a variety of patterns of expression, including a few genes with increased expression in all 7 GI tissues. Other readily identifiable patterns include increased expression in the small and/or large intestine.

Figure 4 shows the results of K-means analysis[5] to cluster gene expression profiles into a total of 30 different classifications. These classifications highlight general patterns of gene expression in the GI tract. Most genes are expressed in an organ-specific (stomach, small intestine, or large intestine) fashion, though this analysis also reveals groups of genes that have highest expression in nonadjacent GI tissues (Figure 4*Z–D′*). In addition, this analysis groups genes whose expression exhibits similar gradients along the A-P axis, such as increasing (Figure 4*P*) or decreasing (Figure 4*L*) within the small intestine. The members of such groups may share mechanisms of gene regulation. These general patterns are also observed for clones that are not relatively highly expressed relative to reference in the various GI segments. Review of "each gene normalized" data for the other 8067 clones represented on the mouse GEM1 microarray identifies an additional 153 clones that have increased expression in the GI tract relative to their median expression in the many mouse tissues tested. K-means analysis shows that these clones also have organ-specific expression profiles in the adult GI tract (data not shown).

Because it is also useful to apply biologically relevant constraints to the analysis of gene expression profiles, an alternative method was used to sort profiles based on anterior to posterior expression within the GI tract (Figure 5*A*). In this figure, expression profiles for each cDNA are read horizontally, with a gray-black scale indicating the expression level. Overall, there is a recognizable pattern of a relatively sharp anterior boundary of expression followed by decreasing expression moving posteriorly. This pattern is evident both with respect to the

level of expression of individual genes, as well as the overall number of genes expressed, and it is most strikingly observed for the large set of genes having anterior-most expression in the duodenum. Most (though not all) of these genes decrease in level of expression moving posteriorly in gut. Similar patterns are observed for genes with anterior-most expression farther along the intestine. Sorting based on posterior to anterior expression or expression from the middle out did not yield such a striking pattern (data not shown). As expected, based on the results of hierarchical clustering (Figure 3), this method shows that there are relatively few genes highly expressed in all 7 GI tissues studied, in the stomach alone, or in all tissues of the small and large intestine.

Functional Classification of Genes Highly Expressed in the Adult Mouse Gastrointestinal Tract

The observation of dynamic and specific regulation of gene expression throughout the adult mouse GI tract suggested that groups of genes encoding proteins with a particular biologic function would be expressed in specific patterns along the A-P axis, and, conversely, that individual regions would preferentially express genes having particular cell functions. To address these possibilities, we classified each of the highly expressed genes based on biochemical function of the encoded protein. Overall, the function of 57% of the genes/ESTs could be identified based on public databases (GenBank and UniGene). The largest fraction of identified genes (20% of the total) were enzymes participating in intermediary metabolism.

To evaluate relationships between functional class and expression patterns, we examined gene expression profiles for each functional category. For example, Figure 5*B* shows the sorted patterns of expression across the A-P axis for genes in each functional category, which reveal several interesting trends. Genes encoding proteins with immunity/defense and structural/cytoskeletal function are generally expressed across the GI tract or at least the small and large intestine. The genes encoding proteins involved in transport functions appear to fall into 3 patterns: expression widely across the GI tract (such as several ion channels and pumps and the polymeric immunoglobulin receptor), expression only in the small intestine (such as intestinal fatty acid binding protein and apolipoproteins A-IV, C-II, and C-III), and expression only in the colon (such as aquaporin 8, involved in water transport, and carbonic anhydrases). Note that this analysis, which is based on ratios vs. reference, is particularly useful for genes that are highly expressed across the GI tract. To examine genes expressed in a segmental or organ-specific pattern, "each gene normalized" profiles (as in Figure 4) were classified by function. Figure 6 shows color-coded profiles based on the K-means–clustered expression profiles in Figure 4. This approach shows that genes encoding proteins functioning in intermediary metabolism, transport, and cell-cell communication have the most dynamically regulated and organ-specific expression profiles, whereas genes in the DNA/RNA process, immune/defense, and structural/cytoskeletal functional groups generally have more level expression across the adult mouse GI tract. Strikingly, there are a number of genes encoding proteins involved in cell-cell communication whose expression is significantly higher in the ileum than any other GI segment. Similar to the results described previously for Figure 5*B,* transport genes show broad GI, small intestinal, or co-

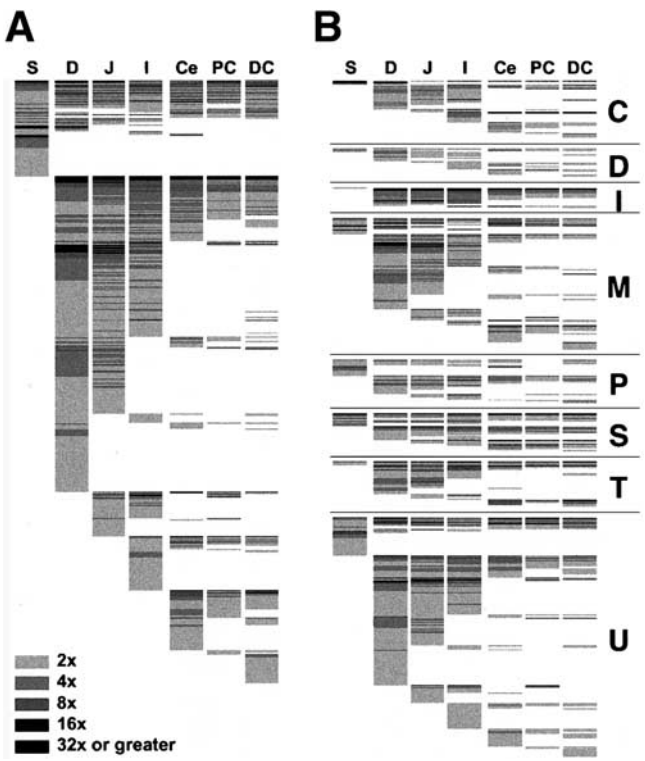

Figure 5. Expression patterns of genes with increased expression in adult mouse GI tissues. (*A*) Levels of expression for highly expressed genes (corresponding to 571 cDNAs with increased expression in at least 1 GI tissue) were defined as $\log_2$ ratios (relative to reference). Expression profiles for each cDNA are read horizontally, with a gray-black shade corresponding to the $\log_2$ ratio for each tissue in which it was expressed 2-fold or more relative to reference. Expression profiles were sorted in Microsoft Excel as described in the Materials and Methods section. (*B*) Gene expression patterns sorted by functional classification, then by profile (as in *A*). Classifications (along vertical axis): C, cell-cell communication; D, DNA/RNA processes; I, immune/defense; M, intermediary metabolism; P, miscellaneous cell processes; S, structure/cytoskeleton; T, transport; U, unknown/other. Tissues (along horizontal axis): S, stomach; D, duodenum; J, jejunum; I, ileum; Ce, cecum; PC, proximal colon; DC, distal colon.

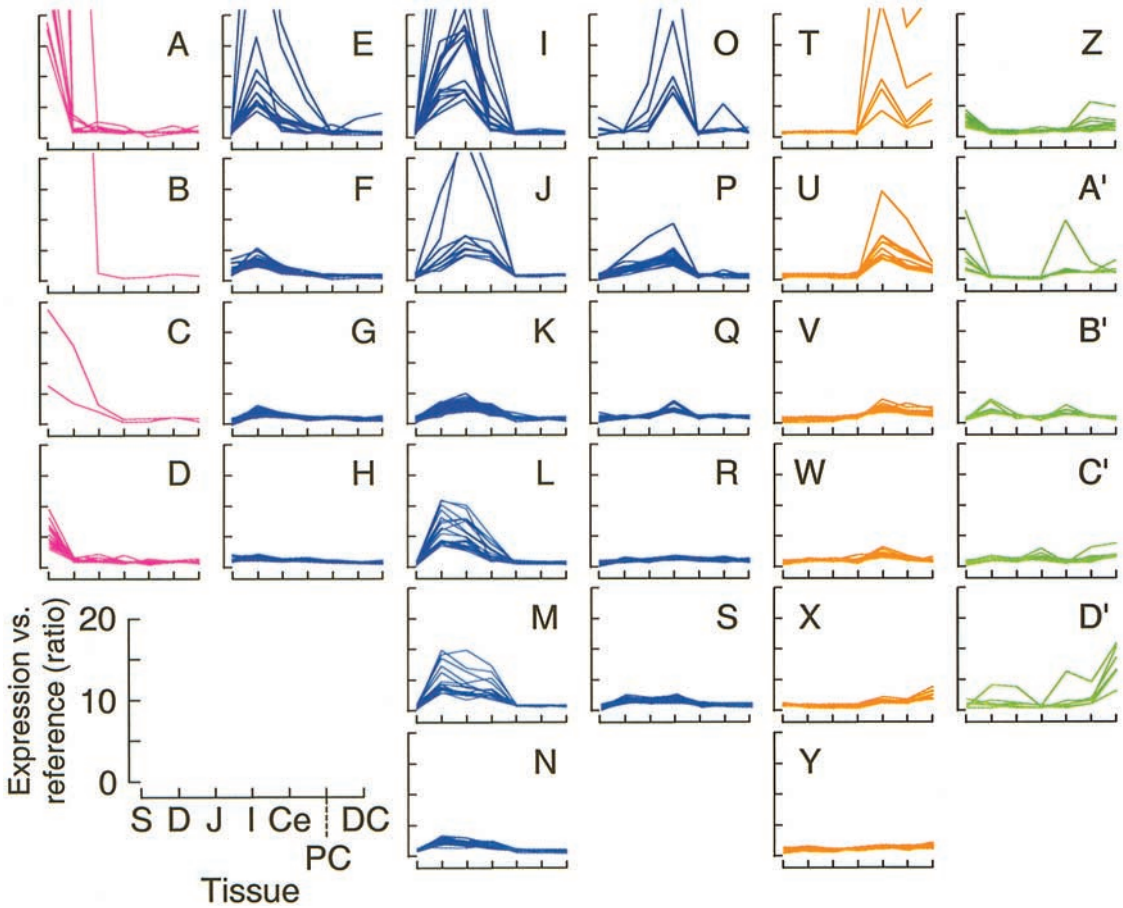

Figure 4.

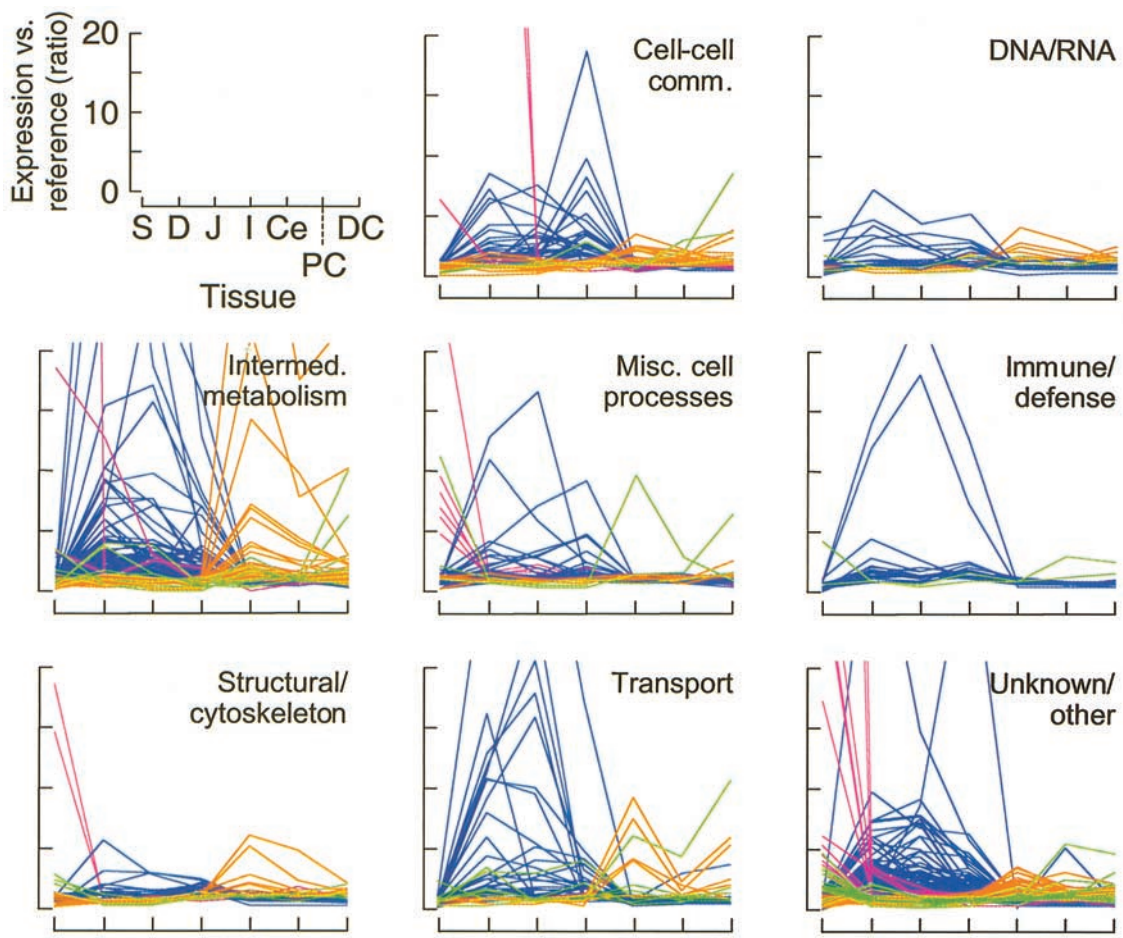

Figure 6.

lonic expression profiles. Within the limitations of the list of genes and splice forms present on the mouse GEM1 microarray, this approach provides a basis for consideration of some of the region-specific gene functions that are strongly regulated with the adult GI tract.

Genes With Large Changes in Expression Between Adjacent Gastrointestinal Segments

Our results described previously using the set of genes that are highly expressed in the adult mouse GI tract suggest that gene expression along the A-P axis follows anatomic divisions (i.e., between stomach and small intestine, and between small intestine and colon). To examine genetically defined boundaries in the adult GI tract using the entire set of 8638 cDNAs represented on the Incyte GEM1 arrays, we identified genes that exhibited large changes in expression between adjacent tissues. This approach allowed the identification of genes that are dynamically regulated within the GI tract but are not highly expressed relative to reference. The locations of significant transitions (greater than 4-fold change in expression between adjacent segments) are shown in Figure 7. Over half of these transitions (205 of 363) occurred between the stomach and duodenum, and most of the rest (100) were observed between ileum and cecum. Interestingly, there were very few changes observed between adjacent small intestinal (duodenum, jejunum, ileum) or large intestinal (cecum, proximal colon, distal colon) tissues. Within the small intestine and colon most changes were decreases from the more anterior tissue to the more posterior (97 decreases in expression out of 156 boundaries, 62%; Figure 7), consistent with the apparent anterior boundaries of expression among the highly expressed genes. Similar results were obtained when smaller changes in expression levels (2-fold or greater) were included in the analysis (data not shown). Overall, these results show a significant similarity in genes expressed within the different regions of small intestine and within the different regions of the large intestine in contrast to the more defined boundaries between stomach/duodenum and ileum/cecum.

Shared Regulatory Elements in Coordinately Regulated GI Genes

The identification of genes displaying coordinate expression along the A-P axis of the adult mouse GI tract suggests that these genes might share regulatory elements or sequences. To identify potential shared gene expression regulatory sequences, we chose genes displaying boundaries of expression between the ileum and cecum but without high expression outside the GI tract, and for which full mouse gene sequence was available in the Celera mouse genome database. As described in Materials and Methods, we identified 3 genes with peak expression in ileum (SI group), and 3 with peak expression in cecum (LI group) (Figure 8*A*). Although these genes were not chosen a priori based on their identities, we found that they included several with well-characterized GI functions, including intestinal fatty acid binding protein, cubilin (the intrinsic factor–cobalamin receptor), and guanylin. Within each group, we found 500–600 bp sequences within 2 kilobases of the 5′ end of the first exon of each gene that contains a shared list of consensus *cis*-regulatory/transcription factor binding elements (Figures 8*B–D*). For each group, a list of 8–9 different elements was found; 4 were shared between the 2 groups. These results suggest that some elements may be important for general intestinal gene expression, whereas others may help to define regional specificity of expression. Interestingly, these putative regulatory sequences include consensus binding sites for transcription factors known to be important for gene regulation in the GI tract, including Cdx2, hepatocyte nuclear factor-1, and forkhead and GATA factors.[12] Many of these elements are found in phylogenetically conserved regions of the genes' 5′ regions (Figures 8*C* and *D*) or in regions that have been previously shown to be important for

Figure 4. K-means classification of gene expression profiles in adult mouse GI tissues. Analysis was performed with the "each gene normalized" expression ratios, using GeneSpring 4.0 to specify a total of 30 different classifications. This number of classifications was suggested by the types of A-P expression profiles apparent in the hierarchical tree analyses (Figure 3). No gene profiles were left unclassified. We found that the use of fewer classifications resulted in more heterogeneous groupings of expression profiles (data not shown). K-means classified profiles are grouped in the figure and colored according to shared features of the profiles: *A–D* (pink), peak expression in stomach; *E–H* (blue), peak expression in duodenum; *I–K* (blue), highest expression in jejunum; *L–N* (blue), decreasing A-P gradient of expression across the small intestine; *O–S* (blue), peak expression in ileum; *T–Y* (orange), peak expression in large intestine; *Z–D′* (green), peaks of expression in more than 1 nonadjacent tissue. The number of cDNAs in each classification were: *A,* 11; *B,* 1; *C,* 2; *D,* 12; *E,* 14; *F,* 18; *G,* 21; *H,* 49; *I,* 18; *J,* 9; *K,* 33; *L,* 22; *M,* 18; *N,* 46; *O,* 6; *P,* 17; *Q,* 16; *R,* 38; *S,* 58; *T,* 5; *U,* 14; *V,* 23; *W,* 20; *X,* 10; *Y,* 43; *Z,* 18; *A′,* 7; *B′,* 7; *C′,* 9; *D′,* 6. S, stomach; D, duodenum; J, jejunum; I, ileum; Ce, cecum; PC, proximal colon; DC, distal colon.

Figure 6. Expression profiles of genes encoding proteins of assigned functional categories. "Each gene normalized" profiles were sorted by functional category of the encoded protein. The color used for each individual profile is the same as in Figure 4 (pink, peak expression in stomach; blue, peak expression in small intestine; orange, peak expression in colon; green, peaks of expression in more than 1 nonadjacent tissue). S, stomach; D, duodenum; J, jejunum; I, ileum; Ce, cecum; PC, proximal colon; DC, distal colon.

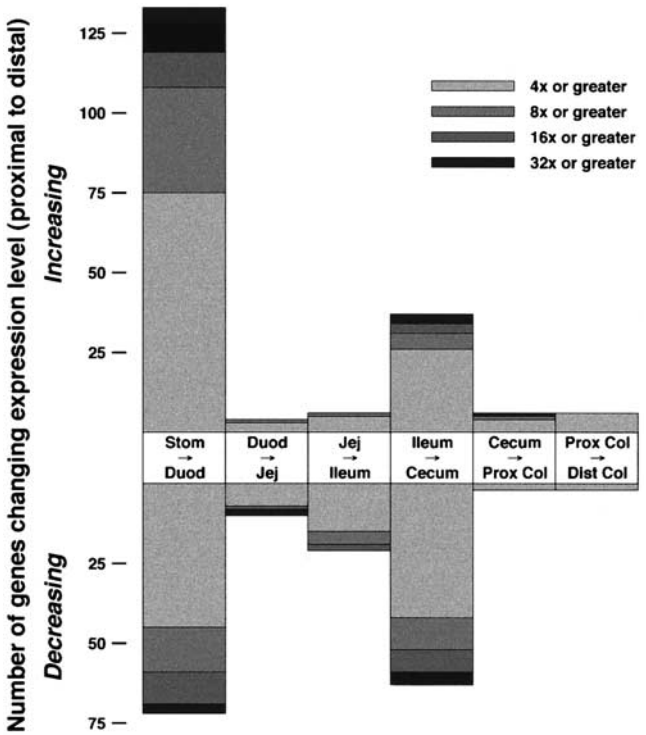

Figure 7. Locations of large changes in gene expression along the adult mouse GI A-P axis. Duplicate expression levels (log_2 ratio) for each cDNA represented on the mouse GEM1 microarray in each tissue were averaged, and differences between adjacent tissues were calculated. The number of cDNAs with increases or decreases of indicated magnitude between adjacent tissues are shown. The expression of 273 cDNAs changed 4-fold or more between adjacent GI segments, including 65 not identified as highly expressed in adult GI tissues (i.e., not on the list of 571 highly expressed adult genes). Because some of these genes had more than one 4-fold change in expression in adjacent tissues, there were a total of 363 transitions identified. Stom, stomach; Duod, duodenum; Jej, jejunum; Prox Col, proximal colon; Dist Col, distal colon.

segment-specific gene expression (Figure 8*C*).[13] Individually, or more likely in combination, these elements represent candidate regulatory sequences for gene expression along the A-P axis of the adult mouse GI tract.

Gene Discovery in the Gastrointestinal Tract

The use of cDNA microarrays containing representations of unknown clones also provides the opportunity for gene discovery relevant to the GI tract. A total of 218 cDNAs and ESTs of the 571 whose corresponding genes are highly expressed in the adult mouse GI tract could not be associated with reported gene sequences in current UniGene clusters or by BLAST searches. These clones exhibit a variety of GI and non-GI expression profiles. We selected 20 of these 218 unidentified clones, representing various expression profiles, for further analysis by comparison of their sequences with the Celera mouse genome database (release 12) from Celera Genomics. Eight of the clones were not highly expressed in any non-GI tissue (Figure 9*A*). All 20 clones exactly matched mouse genomic sequences, and most matched in or near known or predicted genes. In one case (GenBank accession number AA245029), 2 segments of the EST matched 2 genomic segments in opposite order 5 kilobase apart, suggesting an artifact in the EST clone or in the mouse genome assembly. Only 2 of the ESTs could not be associated with known or Celera-predicted mouse genes. In one case, the EST matched both publicly accessible, high-throughput mouse genomic sequence (which allowed mapping to chromosome 14) and unplaced Celera mouse genomic sequence; in the other case, the identified sequence was 34 kilobase from the nearest predicted gene. The sequences of the other 18 matched sequences on or near known or predicted genes, many in the 3′ untranslated region. Eight clones could be associated with known genes because of their proximity within the Celera mouse genome assembly. An additional 6 clones had sequence matching new genes that are homologous to known or predicted genes. The predicted protein products of these genes bear homology to members of protein families that participate in signal transduction, intermediary metabolism, or posttranslational modification of proteins. Thus, many of the unknown ESTs represented on the Mouse GEM1 cDNA microarray correspond to novel genes that are likely important for GI function.

Discussion

The molecular basis of regional specification of gene expression remains a critical problem in GI biology. Although studies of specific genes have shown a number of key mechanisms controlling regional gene expression, the advent of microarray technology provides the opportunity to establish the basis for genome-level studies of regional gene expression in the GI tract. This technology, which allows simultaneous analysis of the expression of thousands of genes, generates information about overall patterns of gene expression in tissues/cells of various sources or with various perturbations, and for gene discovery. To date, most groups have used cDNA microarrays as a tool for the analysis of changes in gene expression in response to perturbations such as small bowel adaptation after intestinal resection,[14] aging,[15] bacterial colonization,[16] inflammation,[17–19] or malignancy.[20–23] This technology has also been used to study pathways of differentiation[24] and apoptosis[25] in colonic cell lines. Microarrays have been used to analyze gene expression profiles among normal tissues from various organs,[26] but

this group did not perform a systematic analysis of segmental expression in the GI tract. We have undertaken an analysis of gene expression patterns across the normal adult GI tract to identify functional relationships and properties of GI segments, as well as to identify novel genes and genes not previously recognized as having a role in the GI tract.

We obtained highly complementary insights into the patterns of highly expressed genes in the adult mouse GI tract by the use of 3 analytical approaches. These approaches differed in the use of different normalizations, data displays, and biologic constraints. The use of expression levels relative to the whole P1 mouse reference was particularly useful for the identification of genes that were highly expressed across the GI tract. Many of these genes are expressed in a GI-specific fashion, as corroborated by comparison of expression profiles examined over a much larger series of microarrays from other non-GI tissues. The use of "each gene normalization" allowed identification of groups of genes, so-called *synexpression groups,*[27] coordinately regulated along the A-P axis of the GI tract. This tissue-specific or organ-specific gene expression may suggest candidates for associations with diseases affecting specific regions of the GI tract.

The identification of coordinately regulated genes, such as shown in Figure 4, suggests similarities in mechanisms of transcriptional regulation that help to define a tissue functionally. Such groups of genes with similar expression profiles, in conjunction with genome sequence data now available, can be used to explore potential mechanisms for segmental regulation of GI gene expression. We found a number of consensus *cis*-regulatory/transcription factor binding elements that are shared within 500–600 bp of the 5′ regions of coordinately regulated genes, including several that have previously been shown to be important for regulation of gene expression in the GI tract (Figure 8). These elements represent candidate regulatory sequences, some of which may function in a tightly regulated manner to direct segmental gene expression in the adult mouse GI tract. Our data, including the identification of elements not well characterized in the GI tract, provide a preliminary step for further large-scale and long-term analysis using biochemical and in silico approaches to the regulation of gene expression in GI tissues.

For most of our analyses, we focused on a set of 571 cDNAs and ESTs that are highly expressed in the adult mouse GI tract relative to reference. This list excludes not only housekeeping genes expressed in all cell types but also genes that may be highly expressed in whole P1 mouse as well as the adult GI tract. However, analysis of the expression profiles of the other clones not on this list yielded patterns of segmental and organ-specific expression that were similar to those observed for the 571 highly expressed clones. Thus, within the complete set of genes and splice variants represented on the mouse GEM1 microarray (approximately 20%–25% of the genes expressed in the mouse genome), our approach shows highly regulated patterns of gene expression along the A-P axis of the adult GI tract.

A common finding with all 3 of our analytical approaches is that the expression patterns of genes in the GI tract generally display distinct boundaries of expression with a variable decrease in expression moving away from this boundary. Most often, there is an anterior sharp boundary of expression with a variable decrease moving posteriorly. The mechanism governing such boundaries of expression is not clear, but it could be mediated in part by Hox transcription factors. The members of this family exhibit overlapping domains of expression with both anterior and posterior boundaries of expression,[28–30] similar to the general patterns of gene expression that we observed. In addition, mutations of Hox genes have significant consequences for gut development.[31–34] Our data support the idea of a Hox code of the GI tract[29,30] that controls regionalization of GI gene expression. This model would predict that a more detailed analysis of GI tracts in Hox gene mutant mice would show alterations of GI anatomy and gene expression patterns.

Gastrointestinal gene expression is regulated not only on a regional basis but also by environmental factors. For example, we found a number of genes highly expressed in the ileum that have also been found to be up-regulated in the distal small intestine after bacterial colonization (information referred to is in the supplement to Hooper et al.[16]). In addition, 1 gene that is down-regulated in the distal small intestine after bacterial colonization (glutathione S-transferase α4, represented on the mouse GEM1 microarray by GenBank number W54349) is highly expressed in stomach, duodenum, and jejunum but not ileum. These results suggest that differential mechanisms or differential exposure to bacteria regulate this gene's level of expression in particular intestinal segments.

Our microarray method and data are validated by comparison of expression profiles of genes whose expression has been previously determined along the A-P axis of the GI tract (Figure 2). For example, the gene encoding the peptide guanylin has been previously shown by northern blot analysis to display segmental expression in the mouse GI tract, with an increase in expression from proximal jejunum to cecum, continued high expression

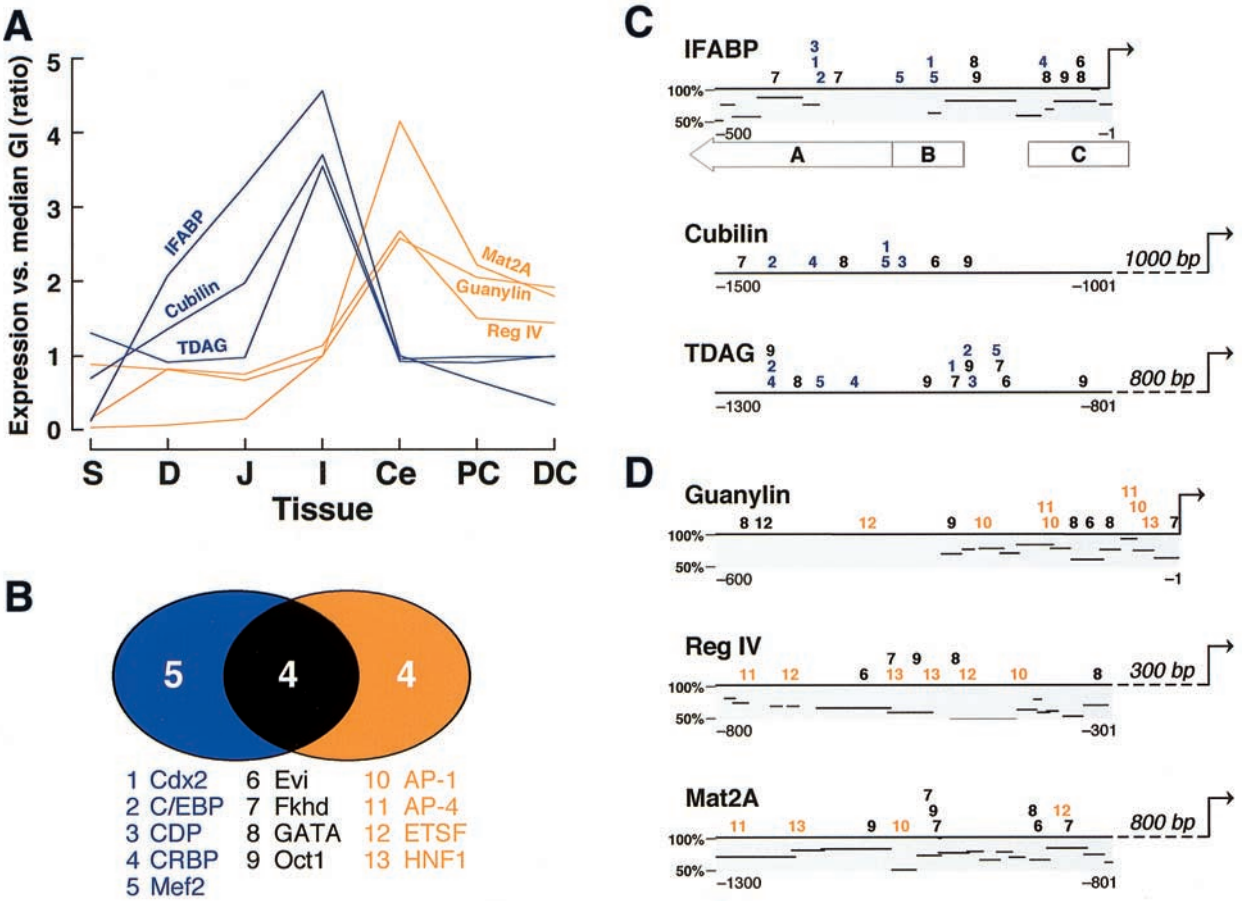

Figure 8.

Incyte Identification	Accession numbers: Incyte	Accession numbers: GenBank	Expression pattern: S D J I Ce PC DC	Location of non-GI expression	Mouse Chr	Celera gene	Genomic location of EST sequence	Identification of encoded protein
ESTs, Weakly similar to JL0144 IL-6 receptor precursor [M.musculus]	313529	W10596		None	7	mCG8205	3' untranslated region	**Prolylcarboxypeptidase (angiotensinase C)**
ESTs	677592	AA212893		Liver	12	– –	Unclear transcript whose gene is ~50 kb downstream of an unknown/unclear gene	No similarities in GenBank or Celera human or mouse databases
ESTs	639994	AA197601		None	1	mCG13077	EST contains a 5' end	No similarities in GenBank or Celera human or mouse databases
ESTs	644941	AA210377		None	17	mCG11771	Alternate or extended 3' untranslated region	New member of the glutathione S-transferase protein family
ESTs	477158	AI510020		Brain, bladder	9	mCG4016	3' untranslated region	**Secretory carrier membrane protein 5**
ESTs	693560	AA239254		Hair (anagen & catagen), epididymis (head), synovium, dorsal root	3	mCG5273	Alternate 3' exon	New member of the long chain fatty acid synthetase protein family
RIKEN cDNA	522713	AA087673		Epididymis	14	Sequence present in Celera small mouse genome fragments only; chromosomal assignment based on public high throughput mouse genome sequence match		
RIKEN cDNA	619728	AA175984		Uterus (estrus)	6	mCG49906	Alternate or extended 3' exon	New member of the adenosyl homocysteinase protein family
DNA segment, Chr 5, Wayne State University 111, expressed	697416	AA245029		Uterus (estrus & pregnant), spleen, lymph node, epididymis (head & tail)	5	mCG1085	5 kb 3' of the predicted Celera first exon	**Onzin**, a 112 amino acid protein induced in uterus by leukemia inhibitory factor (unpublished)
ESTs	719965	AA255150		Hair (anagen), uterus (pregnant, estrus)	18	mCG13251	Alternate or extended 3' untranslated region	**Desmocollin-2**
RIKEN cDNA	574395	AA119577		None	12	mCG54550	Sequence within a Celera-predicted gene but not in the predicted coding region	No similarities in GenBank or Celera human or mouse databases
RIKEN cDNA	459511	AA027452		Pancreas, liver, lactating mammary gland, lymph node	6	mCG16279	Central section of a Celera-predicted transcript containing 3 exons	Member of an uncharacterized new protein family having casein kinase 2 phosphorylation sites and a myristoylation site
ESTs	738239	AA271106		None	2	mCG2101	Actual or alternate 3' untranslated region	Member of an uncharacterized new protein family having a proteinase-associated domain and a proline-rich domain
Mouse Ras association domain family 3 protein mRNA	949241	AA544904		None	10	mCG49016	Alternate or extended 3' untranslated region of Rassf3 (AF332864) not in GenBank	**Ras association domain family 3 (Rassf3)**
RIKEN cDNA	572422	AA105830		Uterus, bladder	2	mCG17528	Alternate splice form	**Major epididymis-specific protein E4**
RIKEN cDNA	484485	AA073913		None	12	mCG14941	5' sequences of the 12 exon gene	**Amnionless**
RIKEN cDNA	598411	AA155071		Olfactory epithelium, nasal epithelium	2	mCG19374	5 kb downstream from a Celera-predicted unknown 6 exon gene spanning 9 kb	No similarities in GenBank or Celera human or mouse databases
ESTs	401906	W82161		Liver	11	mCG17300	5' end of a Celera-predicted transcript	No similarities in GenBank or Celera human or mouse databases
Public domain EST	679920	AA237592		Liver	19	mCG18913	3' untranslated region	**Multidrug resistance associated protein 2 (Mrp2)/canalicular multispecific organic anion transporter**
RIKEN cDNA	420204	W89845		None	5	mCG3485	5' end of the protein-coding region	New member of the peptide aspartate beta hydoxylase protein family

Figure 9.

Figure 8. Consensus *cis*-regulatory/transcription factor binding elements present in the 5′ regions of coordinately regulated intestinal genes. (*A*) Expression profiles of 6 genes with a sharp boundary of expression between the ileum and cecum. cDNAs with highest expression in the ileum (SI group) are shown in blue: AA245078, intestinal fatty acid binding protein (IFABP); AA471960, cubilin; and AA239282, T-cell death associated gene (TDAG). cDNAs with highest expression in the cecum (LI group) are shown in orange: AA498312, methionine adenosyltransferase 2A subunit (Mat2A), AA498457, guanylin, and W18397, Reg IV (a member of the calcium-dependent lectin superfamily).[11] (*B*) Shared 5′ *cis*-regulatory/transcription factor–binding elements in the 5′ regions. Elements present only in the SI group (n = 5) or LI group (n = 4) are shown in blue and orange, respectively; those shared between the groups (n = 4) are shown in black. C/EBP, CCAAT/enhancer-binding protein; CDP, CCAAT-displacement protein; CRBP, CReb–binding protein family; Fkhd, forkhead/winged helix domain factors; ETSF, ETS family; HNF-1, hepatocyte nuclear factor-1. (*C* and *D*) Locations of putative *cis*-regulatory/transcription factor–binding elements in 500–600 bp of the 5′ upstream regions of the selected genes. SI genes are shown in (*C*), LI genes are shown in (*D*). Numbers designate elements as in (*B*). Below the schematic 5′ region for 4 of the genes is a gray box with lines designating regions of high-sequence identity (50%–100%) over a window of at least 200 bp between the corresponding regions of the mouse and human gene orthologues. Boxes under the intestinal fatty acid binding protein gene show regions previously shown to be important for regulation of gene expression along the A-P axis[13]: *A,* −1178 to −277 bp; *B,* −277 to −185 bp; *C,* −103 to +28 bp.

in the proximal colon, and a decrease in expression in the distal colon.[35] The expression profile in the present study was similar (Figure 2), as predicted. Another example is adenosine deaminase, which has previously been shown to be expressed in duodenum and jejunum,[36,37] corroborating the pattern observed here for a public domain EST that is identical in sequence to the mouse adenosine deaminase gene. Thus, known patterns of gene expression in the adult GI tract are also observed by expression analysis using cDNA microarrays.

The problem of regional gene expression not only has theoretical ramifications, but also practical ones, in that disordered regional specification may contribute to human disease. An example of altered regional specification is the intestinal metaplasia observed in Barrett's esophagus, a premalignant condition. Genes that play a role in malignancy and that we found display strong regional expression patterns in the GI tract include the transcription factors Cdx2[38,39] and Max (which forms heterodimers with members of the Myc oncogene family[40]).

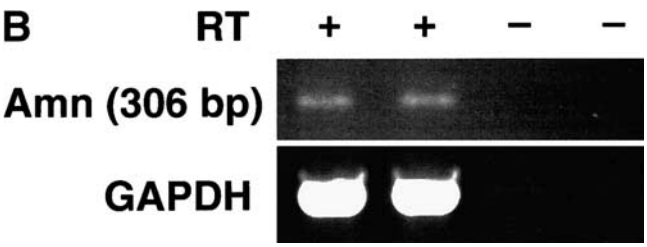

Figure 9. Gene discovery using unknown/uncharacterized genes that are highly expressed in the adult mouse GI tract. (*A*) A total of 20 unknown clones that could not be characterized using UniGene and BLAST searches were selected based on interesting GI and non-GI expression patterns. The list is ordered based on hierarchical clustering of the expression profiles of these clones in the GI tract, using $\log_2$-transformed ratios relative to reference. GI expression profiles are shown by colored squares, with 2 squares (1 for each microarray hybridization) for each tissue. Colors are graded with red indicating increased expression and green indicating low expression relative to reference. Sequences of the clones were compared with the mouse genome database of Celera Genomes as described in Materials and Methods. The location of expression (greater than 2-fold above reference) in normal adult non-GI tissues in our mouse gene expression database, location of the clone sequence relative to the gene, and the identified or putative gene or gene family are indicated. S, stomach; D, duodenum; J, jejunum; I, ileum; Ce, cecum; PC, proximal colon; DC, distal colon; Chr, chromosome. (*B*) Reverse transcription–polymerase chain reaction analysis of *amnionless* gene expression in adult mouse ileum. cDNA was prepared from poly A^+ RNA obtained from adult ileum as described and used for polymerase chain reaction to detect expression of *amnionless* and glyceraldehyde-3-phosphatedehydrogenase (control) as described in Materials and Methods.

Among the most highly expressed genes in the small intestine and colon relative to the whole P1 mouse reference are those encoding proteins with immune function, including immunoglobulin A heavy chain (which is represented with 3 different probes on the GEM1 microarray), the polymeric immunoglobulin receptor (which transcytoses immunoglobulin A across the intestinal epithelium into the GI lumen), immunoglobulin G, a variety of H-2 major histocompatibility antigens, and β_2-microglobulin. These findings point out the richness of the intestine as an immune organ, with function developing after birth as the GI lumen is in topologic continuity with the external environment.

Our results also show the power of our microarray approach for gene discovery with respect to the GI tract. In particular, the combination of previously isolated EST clones, gene expression profiles, and genome sequence data and annotation allows the identification of genes, including new members of gene families, not previously known to be expressed in the adult GI tract as well as new genes not previously characterized. The use of bioinformatics to study genome sequence data can be used also to assign ESTs containing sequences of 3′ untranslated regions to specific genes and to identify gene and protein families not previously recognized. The expression data can be used to aid and prioritize the gene discovery process by pointing to systems and functions in which the genes and their encoded proteins participate.

One clone exhibiting a particularly interesting expression profile corresponds to the mouse *amnionless (Amn)* gene. Microarray analysis (Figure 9*A*), confirmed by reverse transcription–polymerase chain reaction (Figure 9*B*), showed that *Amn* is highly expressed in the small intestine, with an increasing gradient moving from anterior to posterior, with highest expression in the ileum, but it is not highly expressed in any of the adult non-GI tissues studied. Mutation of this gene was recently shown[41] to be responsible for the gastrulation defect in a transgene-induced insertional mutant mouse line.[42] The encoded protein contains a transmembrane-spanning domain and a cysteine-rich domain similar to those found in a variety of regulators of bone morphogenetic proteins, which play important roles in the development and function of a variety of organ systems.[43] For example, bone morphogenetic protein-2 and bone morphogenetic protein-4 are known to play important roles in gastrointestinal development.[2,44] Kalantry et al.[41] postulate that, as a membrane-bound protein, Amn might modulate bone morphogenetic protein–receptor function. A potential role for Amn in the adult intestine is unknown, but it could play a role in cell-cell interactions important for intestinal mucosal integrity.

An example of a gene identified based on an EST containing sequence 3′ to a known gene (that might represent an alternate or extended 3′ region) is that encoding desmocollin 2. Desmocollin 2 is 1 of 3 desmocollins, cadherin family members that are expressed in desmosomes. At least 2 alternately spliced transcripts from the mouse desmocollin 2 gene are known.[45] Unlike Desmocollin 1 and Desmocollin 3, whose expression is limited to epidermis, desmocollin 2 is expressed in a wide variety of tissues, including the GI tract.[45,46] Our data in mouse are similar, showing expression widely in the GI tract as well as in hair and uterus (Figure 9*A*). Thus, our approach can be useful for the identification of splice variants and their expression patterns.

In summary, the anatomically recognized organs of the GI tract (stomach, small intestine, colon) can be defined by a genome-level analysis of gene expression profiles. However, at a molecular level the distinctions between various regions of the small intestine and colon are less striking. A wide variety of proteins with various physiologic functions are subject to segmental regulation within the GI tract. The further identification and study of genes coordinately regulated in a segmental fashion along the A-P axis will likely provide new understanding of the development, function, and pathophysiology of the GI tract.

References

1. Simon TC, Gordon JI. Intestinal epithelial cell differentiation: new insights from mice, flies and nematodes. Curr Opin Genet Dev 1995;5:577–586.
2. Roberts DJ. Molecular mechanisms of development of the gastrointestinal tract. Dev Dyn 2000;219:109–120.
3. Bowtell DDL. Options available—from start to finish—for obtaining expression data by microarray. Nat Genet 1999;21(Suppl): 25–32.
4. Eisen MB, Spellman PT, Brown PO, Botstein D. Cluster analysis and display of genome-wide expression patterns. Proc Natl Acad Sci U S A 1998;95:14863–14868.
5. Gordon AD. Classification (2nd ed.). Monographs on Statistics and Applied Probability, 82. Boca Raton, FL: Chapman & Hall/ CRC Press, 1999:41–48.
6. International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. Nature 2001;409: 860–921.
7. Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, Smith HO, Yandell M, Evans CA, Holt RA, Gocayne JD, Amanatides P, Ballew RM, Huson DH, Wortman JR, Zhang Q, Kodira CD, Zheng XH, Chen L, Skupski M, Subramanian G, Thomas PD, Zhang J, Gabor Miklos GL, Nelson C, Broder S, Clark AG, Nadeau J, McKusick VA, Zinder N, Levine AJ, Roberts RJ, Simon M, Slayman C, Hunkapiller M, Bolanos R, Delcher A, Dew I, Fasulo D, Flanigan M, Florea L, Halpern A, Hannenhalli S, Kravitz S, Levy S, Mobarry C, Reinert K, Remington K, Abu-Threideh J, Beasley E, Biddick K, Bonazzi V, Brandon R, Cargill M, Chandramouliswaran I, Charlab R, Chaturvedi K, Deng Z, Di Francesco V, Dunn P, Eilbeck K, Evangelista C, Gabrielian AE, Gan W, Ge W, Gong F, Gu Z, Guan P, Heiman TJ, Higgins ME, Ji RR, Ke Z, Ketchum KA, Lai Z, Lei Y, Li Z, Li J, Liang Y, Lin X, Lu F, Merkulov GV, Milshina N, Moore HM, Naik AK, Narayan VA, Neelam B, Nusskern D, Rusch DB, Salzberg S, Shao W, Shue B, Sun J, Wang Z, Wang A, Wang X, Wang J, Wei M, Wides R, Xiao C, Yan C, Yao A, Ye J, Zhan M, Zhang W, Zhang H, Zhao Q, Zheng L, Zhong F, Zhong W, Zhu S, Zhao S, Gilbert D, Baumhueter S, Spier G, Carter C, Cravchik A, Woodage T, Ali F, An H, Awe A, Baldwin D, Baden H, Barnstead M, Barrow I, Beeson K, Busam D, Carver A, Center A, Cheng ML, Curry L, Danaher S, Davenport L, Desilets R, Dietz S, Dodson K, Doup L, Ferriera S, Garg N, Gluecksmann A, Hart B, Haynes J, Haynes C, Heiner C, Hladun S, Hostin D, Houck J, Howland T, Ibegwam C, Johnson J, Kalush F, Kline L, Koduru S, Love A, Mann F, May D, McCawley S, McIntosh T, McMullen I, Moy M, Moy L, Murphy B, Nelson K, Pfannkoch C, Pratts E, Puri V, Qureshi H, Reardon M, Rodriguez R, Rogers YH, Romblad D, Ruhfel B, Scott R, Sitter C, Smallwood M, Stewart E, Strong R, Suh E, Thomas R, Tint NN, Tse S, Vech C, Wang G, Wetter J, Williams S, Williams M, Windsor S, Winn-Deen E, Wolfe K, Zaveri J, Zaveri K, Abril JF, Guigo R, Campbell MJ, Sjolander KV, Karlak B, Kejariwal A, Mi H, Lazareva B, Hatton T, Narechania A, Diemer K, Muruganujan A, Guo N, Sato S, Bafna V, Istrail S, Lippert R, Schwartz R, Walenz B, Yooseph S, Allen D, Basu A, Baxendale J, Blick L, Caminha M, Carnes-Stine J, Caulk P, Chiang YH, Coyne M, Dahlke C, Mays A, Dombroski M, Donnelly M, Ely D, Esparham S, Fosler C, Gire H, Glanowski S, Glasser K, Glodek A, Gorokhov M, Graham K, Gropman B, Harris M, Heil J, Henderson S, Hoover J, Jennings D, Jordan C, Jordan J, Kasha J, Kagan L, Kraft C, Levitsky A, Lewis M, Liu X, Lopez J, Ma D, Majoros W, McDaniel J, Murphy S, Newman M, Nguyen T, Nguyen N, Nodell M, Pan S, Peck J, Peterson M, Rowe W, Sanders R, Scott J, Simpson M, Smith T, Sprague A, Stockwell T, Turner R, Venter E, Wang M, Wen M, Wu D, Wu M, Xia A, Zandieh A, Zhu X. The sequence of the human genome. Science 2001;291:1304–1351.
8. Schwartz S, Zhang Z, Frazer KA, Smit A, Riemer C, Bouck J, Gibbs

R, Hardison R, Miller W. PipMaker—a web server for aligning two genomic DNA sequences. Genome Res 2000;10:577–586.
9. Wingender E, Chen X, Hehl R, Karas H, Liebich I, Matys V, Meinhardt T, Prüß M, Reuter I, Schacherer F. TRANSFAC: an integrated system for gene expression regulation. Nucleic Acids Res 2000;28:316–319.
10. Thim L, Mørtz E. Isolation and characterization of putative trefoil peptide receptors. Regul Pept 2000;90:61–68.
11. Hartupee JC, Zhang H, Bonaldo MF, Soares MB, Dieckgraefe BK. Isolation and characterization of a cDNA encoding a novel member of the human regenerating protein family: Reg IV. Biochim Biophys Acta 2001;1518:287–293.
12. Wiginton DA. Gene regulation: the key to intestinal development. In: Sanderson IR, Walker WA, eds. Development of the gastrointestinal tract. Hamilton, Ontario: B.C. Decker, 1999:13–36.
13. Cohn SM, Simon TC, Roth KA, Birkenmeier EH, Gordon JI. Use of transgenic mice to map *cis*-acting elements in the intestinal fatty acid binding protein gene (*Fabpi*) that control its cell lineage-specific and regional patterns of expression along the duodenal-colonic and crypt-villus axes of the gut epithelium. J Cell Biol 1992;119:27–44.
14. Stern LE, Erwin CR, Falcone RA, Huang FS, Kemp CJ, Williams JL, Warner BW. cDNA microarray analysis of adapting small bowel after intestinal resection. J Pediatr Surg 2001;36:190–195.
15. Lee HM, Greenley GH, Englander EW. Age-associated changes in gene expression patterns in the duodenum and colon of rats. Mech Ageing Dev 2001;122:355–371.
16. Hooper LV, Wong MH, Thelin A, Hansson L, Falk PG, Gordon JI. Molecular analysis of commensal host-microbial relationships in the intestine. Science 2001;291:881–884.
17. Dieckgraefe BK, Stenson WF, Korzenik JR, Swanson PE, Harrington CA. Analysis of mucosal gene expression in inflammatory bowel disease by parallel oligonucleotide arrays. Physiol Genomics 2000;4:1–11.
18. Honda M, Kaneko S, Kawai H, Shirota Y, Kobayashi K. Differential gene expression between chronic hepatitis B and C hepatic lesion. Gastroenterology 2001;120:955–966.
19. Lawrance IC, Fiocchi C, Chakravarti S. Ulcerative colitis and Crohn's disease: distinctive gene expression profiles and novel susceptibility candidate genes. Hum Mol Genet 2001;10:445–456.
20. Alon U, Barkai N, Notterman DA, Gish K, Ybarra S, Mack D, Levine AJ. 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;96:6745–6750.
21. Schraml P, Kononen J, Bubendorf L, Moch H, Bissig H, Nocito A, Mihatsch MJ, Kallioniemi OP, Sauter G. Tissue microarrays for gene amplification surveys in many different tumor types. Clin Cancer Res 1999;5:1966–1975.
22. Getz G, Levine E, Domany E. Coupled two-way clustering analysis of gene microarray data. Proc Natl Acad Sci U S A 2000;97:12079–12084.
23. Hippo Y, Yashiro M, Ishii M, Taniguchi H, Tsutsumi S, Hirakawa K, Kodama T, Aburatani H. Differential gene expression profiles of scirrhous gastric cancer cells with high metastatic potential to peritoneum or lymph nodes. Cancer Res 2001;61:889–895.
24. Mariadason JM, Corner GA, Augenlicht LH. Genetic reprogramming in pathways of colonic cell maturation induced by short chain fatty acids: comparison with trichostatin A, sulindac, and curcumin and implications for chemoprevention of colon cancer. Cancer Res 2000;60:4561–4572.
25. Manos EJ, Jones DA. Assessment of tumor necrosis factor receptor and Fas signaling pathways by transcriptional profiling. Cancer Res 2001;61:433–438.
26. Miki R, Kadota K, Bono H, Mizuno Y, Tomaru Y, Carninci P, Itoh M, Shibata K, Kawai J, Konno H, Watanabe S, Sato K, Tokusumi Y, Kikuchi N, Ishii Y, Hamaguchi Y, Nishizuka I, Goto H, Nitanda H, Satomi S, Yoshiki A, Kusakabe M, DeRisi JL, Eisen MB, Iyer VR, Brown PO, Muramatsu M, Shimada H, Okazaki Y, Hayashizaki Y. Delineating developmental and metabolic pathways in vivo by expression profiling using the RIKEN set of 18,816 full-length enriched mouse cDNA arrays. Proc Natl Acad Sci U S A 2001;98:2199–2204.
27. Niehrs C, Pollet N. Synexpression groups in eukaryotes. Nature 1999;402:483–487.
28. Yokouchi Y, Sakiyama J, Kuroiwa A. Coordinated expression of *Abd-B* subfamily genes of the *HoxA* cluster in the developing digestive tract of chick embryo. Dev Biol 1995;169:76–89.
29. Sekimoto T, Yoshinobu K, Yoshida M, Kuratani S, Fujimoto S, Araki M, Tajima N, Araki K, Yamamura K. Region-specific expression of murine *Hox* genes implies the *Hox* code-mediated patterning of the digestive tract. Genes Cells 1998;3:51–64.
30. Pitera JE, Smith VV, Thorogood P, Milla PJ. Coordinated expression of 3′ *Hox* genes during murine embryonic gut development: an enteric *Hox* code. Gastroenterology 1999;117:1339–1351.
31. Boulet AM, Capecchi MR. Targeted disruption of *hoxc-4* causes esophageal defects and vertebral transformations. Dev Biol 1996;177:232–249.
32. Aubin J, Chailler P, Ménard D, Jeannotte L. Loss of *Hoxa5* gene function in mice perturbs intestinal maturation. Am J Physiol 1999;277:C965–C973.
33. Kondo T, Dollé P, Zákány J, Duboule D. Function of posterior *HoxD* genes in the morphogenesis of the anal sphincter. Development 1996;122:2651–2659.
34. Zákány J, Duboule D. Hox genes and the making of sphincters. Nature 1999;401:761.
35. Whitaker TL, Witte DP, Scott MC, Cohen MB. Uroguanylin and guanylin: distinct but overlapping patterns of messenger RNA expression in mouse intestine. Gastroenterology 1997;113:1000–1006.
36. Dusing M, Brickner A, Thomas M, Wiginton D. Regulation of duodenal specific expression of the human adenosine deaminase gene. J Biol Chem 1997;272:26634–26642.
37. Dusing M, Brickner A, Lowe S, Cohen M, Wiginton D. A duodenum-specific enhancer regulates expression along three axes in the small intestine. Am J Physiol 2000;279:G1080–G1093.
38. Chawengsaksophak K, James R, Hammond VE, Köntgen F, Beck F. Homeosis and intestinal tumours in *Cdx2* mutant mice. Nature 1997;386:84–87.
39. Wicking C, Simms LA, Evans T, Walsh M, Chawengsaksophak K, Beck F, Chenevix-Trench G, Young J, Jass J, Leggett B, Wainwright B. CDX2, a human homologue of Drosophila caudal, is mutated in both alleles in a replication error positive colorectal cancer. Oncogene 1998;17:657–659.
40. Baudino TA, Cleveland JL. The Max network gone mad. Mol Cell Biol 2001;21:691–702.
41. Kalantry S, Manning S, Haub O, Tomihara-Newberger C, Lee H-G, Fangman J, Disteche CM, Manova K, Lacy E. The amnionless gene, essential for mouse gastrulation, encodes a visceral-endoderm–specific protein with an extracellular cysteine-rich domain. Nat Genet 2001;27:412–416.
42. Tomihara-Newberger C, Haub O, Lee H-G, Soares V, Manova K, Lacy E. The *amn* gene product is required in extraembryonic tissues for the generation of middle primitive streak derivatives. Dev Biol 1998;204:34–54.
43. Ducy P, Karsenty G. The family of bone morphogenetic proteins. Kidney Int 2000;57:2207–2214.
44. Smith DM, Tabin CJ. BMP signalling specifies the pyloric sphincter. Nature 1999;402:748–749.
45. Lorimer JE, Hall LS, Clarke JP, Collins JE, Fleming TP, Garrod DR. Cloning, sequence analysis and expression pattern of mouse desmocollin 2 (DSC2), a cadherin-like adhesion molecule. Mol Membr Biol 1994;11:229–236.

46. Nuber UA, Schäfer S, Schmidt A, Koch PJ, Franke WW. The widespread human desmocollin 2 (Dsc 2) and tissue-specific patterns of synthesis of various desmocollin subtypes. Eur J Cell Biol 1995;66:69–74.

Received September 14, 2001. Accepted January 17, 2002.

Address requests for reprints to: Michael D. Bates, M.D., Ph.D., Division of Pediatric Gastroenterology, Hepatology and Nutrition, Children's Hospital Medical Center, 3333 Burnet Avenue, Cincinnati, Ohio 45229. e-mail: michael.bates@chmcc.org; fax: (513) 636-7805.

Supported by the Howard Hughes Medical Institute; the Children's Hospital Research Foundation; by National Institutes of Health grants: K08 DK 02791 (M.D.B.), R01 DK 52343 (D.W.), R01 DK 53234 (B.W.W.), R01 DK 47318 (M.B.C.), and R01 ES 008822 (B.J.A.); and by the Children's Hospital Campaign for Children Fund (B.W.W.).

The authors thank the many contributors to the Mouse Gene Expression Database at Children's Hospital Medical Center and the University of Cincinnati, Dr. Mario Medvedovic for advice on statistical methods, Dr. Kennan V. Kellaris (Celera Genomics) for advice regarding the Celera Discovery System, and Elizabeth Florence and Steve Wowk for assistance with sample preparation.