Mamm Genome (2009) 20:516–527
DOI 10.1007/s00335-009-9212-7

Biomarkers of human gastrointestinal tract regions

Elena Maria Comelli · Sofiane Lariani · Marie-Camille Zwahlen · Grigorios Fotopoulos · James Anthony Holzwarth · Christine Cherbut · Gian Dorta · Irène Corthésy-Theulaz · Martin Grigorov

Received: 30 April 2009 / Accepted: 23 July 2009 / Published online: 27 August 2009

Abstract Dysregulation of intestinal epithelial cell performance is associated with an array of pathologies whose onset mechanisms are incompletely understood. While whole-genomics approaches have been valuable for studying the molecular basis of several intestinal diseases, a thorough analysis of gene expression along the healthy gastrointestinal tract is still lacking. The aim of this study was to map gene expression in gastrointestinal regions of healthy human adults and to implement a procedure for microarray data analysis that would allow its use as a reference when screening for pathological deviations. We analyzed the gene expression signature of antrum, duodenum, jejunum, ileum, and transverse colon biopsies using a biostatistical method based on a multivariate and univariate approach to identify region-selective genes. One hundred sixty-six genes were found responsible for distinguishing the five regions considered. Nineteen had never been described in the GI tract, including a semaphorin probably implicated in pathogen invasion and six novel genes. Moreover, by crossing these genes with those retrieved from an existing data set of gene expression in the intestine of ulcerative colitis and Crohn's disease patients, we identified genes that might be biomarkers of Crohn's and/or ulcerative colitis in ileum and/or colon. These include *CLCA4* and *SLC26A2*, both implicated in ion transport. This study furnishes the first map of gene expression along the healthy human gastrointestinal tract. Furthermore, the approach implemented here, and validated by retrieving known gene profiles, allowed the identification of promising new leads in both healthy and disease states.

Electronic supplementary material The online version of this article (doi:10.1007/s00335-009-9212-7) contains supplementary material, which is available to authorized users.

E. M. Comelli · G. Fotopoulos · C. Cherbut · I. Corthésy-Theulaz
Department of Nutrition and Health, Nestlé Research Center, Vers chez les Blanc, 1000 Lausanne 26, Switzerland

E. M. Comelli (✉)
Department of Nutritional Sciences, Faculty of Medicine, University of Toronto, 150 College Street, FitzGerald Building, Room 308A, Toronto, ON M5S 3E2, Canada
e-mail: elena.comelli@utoronto.ca

S. Lariani · M.-C. Zwahlen · J. A. Holzwarth · M. Grigorov
Department of BioAnalytical Science, Nestlé Research Center, Vers chez les Blanc, 1000 Lausanne 26, Switzerland

G. Dorta
Department of Gastroenterology, University Hospital CHUV, 1011 Lausanne, Switzerland

Present Address:
G. Fotopoulos
Novartis Consumer Health, 1260 Nyon 1, Switzerland

Present Address:
S. Lariani
Nestlé R&D Centre York, York, UK

Introduction

The gastrointestinal (GI) tract is a complex system characterized by multiple biological functions that are anatomically distributed. Besides being the site for food processing and nutrient absorption, it is also the largest immune organ of the body and, being at the interface with the external environment, constitutes an important barrier against ingested pathogens and toxins. Furthermore, the intestine is the home of a large microbial

community, roughly comprising 10^{14} cells, which is now considered an additional metabolic organ of our body (Backhed et al. 2005). These multiple functions significantly rely on the epithelial cells. Given their ability to integrate a large variety of environmental signals, including those coming from the gut resident microbiota, these cells are the main regulators of intestinal mucosal homeostasis (Clavel and Haller 2007; McCole and Barrett 2007). Hence, dysregulation of intestinal epithelial cell performance is associated with an array of different pathologies. For example, several studies have shown that immune and epithelial cells concur with T-helper cells in the pathophysiology of inflammatory bowel disease (IBD) by undergoing phenotypic modifications that mainly affect the maintenance of the epithelial barrier. These modifications include reduced production of α-defensins (Wehkamp et al. 2005), altered expression of mucin genes (Buisine et al. 1999, 2001), expression of IL-15 and TGFβ (Yoshikai 1999), and dysregulation of the pregnane X receptor target genes (Langmann et al. 2004). Also, it has been suggested that differences exist in the phenotypic features of epithelial cells between ulcerative colitis (UC) and Crohn's disease (Andoh et al. 2001), although the onset mechanisms of these pathologies are incompletely understood.

Global gene expression profile analyses of diseased tissues greatly contributed to the understanding of how deviation from healthy biological pathways may be associated with specific phenotypes. These include neoplastic diseases such as colorectal cancer (Reichling et al. 2005; Sugiyama et al. 2005), celiac disease (Diosdado et al. 2004; Juuti-Uusitalo et al. 2007), and idiopathic diseases such as IBD (Csillag et al. 2007; Langmann et al. 2004; Moehle et al. 2006). The interpretation of a pathology-related gene expression signature relies on knowing the healthy profile. Nonetheless, genomics approaches have seldom been used for the study of the unaffected GI tract (Fleet 2007). Initial studies conducted in rodents furnished a basic understanding of gene expression modulation along the healthy GI tract. In a pioneer work, a gene expression analysis based on a cDNA array representing about 8600 genes showed that biological functions and gene expression are highly correlated along the murine GI tract (Bates et al. 2002). Indeed, while some transcripts are commonly found in different regions, distinct intestinal segments are characterized by specific gene expression profiles (Anderle et al. 2005; Fleet 2007; Mutch et al. 2004). Other studies examined gene expression along the crypt-villus axis in the small intestine and showed that a specific gene expression program characteristic of intestinal epithelial cell maturation also exists (Mariadason et al. 2005; Stegmann et al. 2006). On the other hand, less is known about how gene expression is modulated in the healthy human GI tract. In the large intestine, the existence of transcripts that are specific to its different regions as well as of transcripts that are gradually modulated longitudinally has been shown (Glebov et al. 2003; Lapointe et al. 2008). The regional character of gene expression in the large intestine could be partially traced back to the different embryological origins of its proximal and distal parts (Lapointe et al. 2008), but it has also been suggested to play a role in the susceptibility of the distinct colonic segments to different types of colorectal carcinomas (Glebov et al. 2003).

To our knowledge, gene expression variation along the cephalocaudal axis of the whole human gastrointestinal tract has not been investigated. By analyzing gene expression in biopsies taken from the stomach, duodenum, jejunum, ileum, and transverse colon of healthy adults, this study contributes to our understanding of gene expression evolution along this system. Furthermore, this work addresses how the generated profiles of gene expression can be used as a reference when screening for pathological deviations. The underlying assumption of current microarray studies is the belief that it is, in principle, possible to fully describe a biological sample, a snapshot of a precise physiological state, by a set of measures of gene expression levels. In this context, the levels of mRNA that, for example, Affymetrix probe sets account for can be interpreted as the descriptors of a given physiological state or biological sample. This information is sometimes difficult to exploit for the identification of specific targets or disease biomarkers because of the large amount of noise intrinsic to microarray analyses. Noise in this context can be related to fluctuation due to the experimental conditions, to the presence of outliers, and to genes presenting no modulation between groups.

In this study we produced a map of gene expression in different regions of the healthy human GI tract and implemented a biostatistical method for noise filtering and statistical significance assessment. This method is based on the combination of a multivariate approach, principal component analysis (PCA) for a first data compression and signal extraction, and an univariate approach using robust statistics for the estimation of statistical significance of the selected genes. The built workflow furnishes a flexible but robust way to extract knowledge to recognize significant expression patterns with high confidence within the set of the identified modulated genes.

By first identifying specific biomarkers of healthy human gut segments and then by crossing them with genes mostly affected in IBD we were able to identify both interesting new leads such as novel ileal and colonic genes, and promising candidate biomarkers of Crohn's disease and/or ulcerative colitis in ileum and/or colon.

Materials and methods

Samples collection

Biopsies (10–15 mg) were collected by upper GI endoscopy (with a pediatric colonoscope to reach the proximal jejunum) and colonoscopy from the antrum, duodenum, jejunum, ileum, and transverse colon (hereafter colon) of two male and two female healthy human subjects recruited in accordance with the Helsinki declaration and with previous approval by the Ethical Committee of the Canton de Vaud, Lausanne, Switzerland. The individuals, informed about the scope of the study, were between 20 and 30 years, did not have any gastrointestinal or cardiovascular disorder, did not use any drugs during the 4 weeks before sampling, and were *Helicobacter pylori* negative. Study participants were required to fast for 12 h and were given a bowel preparation (Fordtran's and Cololyt solutions) to prepare the intestine the evening before endoscopy. All endoscopic procedures were performed by a gastroenterologist at the University Hospital of the Canton de Vaud (CHUV), Lausanne, Switzerland. The biopsies, which were expected to mainly contain mucosal tissue, were snap-frozen and stored at –80°C until use (within 2 weeks from sampling).

RNA isolation and gene array hybridization

Total RNA was extracted using the TriPure Isolation reagent (Roche Diagnostics AG, Rotkreuz, Switzerland) and Dnase-I treated and purified with the Nucleospin kit (Macherey-Nagel AG, Oensingen, Switzerland). RNA quality was assessed with the Agilent 2100 Bioanalyzer and the RNA 6000 Nano Chip kit (Agilent Technologies, Waldbronn, Germany). RNA (5 μg) from each sample was then independently amplified, labeled, and hybridized to the HG-U133A Affymetrix chip (Affymetrix, High Wycombe, UK), according to the manufacturer's instructions. The chips were scanned at 488 nm with an argon-ion laser (Agilent Biotechnologies and Affymetrix) and expression signals were generated and normalized using the MAS 5.0 algorithm (Affymetrix). All chips used in this study were processed in line with and passed the quality control steps built in the Affymetrix standard procedure for gene array hybridization, including spiking with a prokaryotic control cocktail to assess hybridization performance. The entire data set has been deposited at the NCBI Gene Expression Omnibus (GEO) repository (http://www.ncbi.nlm.nih.gov/geo/) under accession number GSE10867. Gene expression of the samples was not investigated with any additional technique since microarray measurements have been shown to be reproducible across different platforms and therefore they can be used independently (Chou et al. 2008; Shi et al. 2006).

Identification of region-selective biomarkers

After a preprocessing step, including normalization (MAS 5.0, Affymetrix) and log transformation of the hybridization signals, the data were organized into a matrix of 20 rows and 22,283 columns, corresponding to the number of arrays (each representing a biological sample of one of the five gut compartments labeled according to the gender of the corresponding individual) and the number of probe sets (variables), respectively. To check for intrinsic clustering among the 20 biological samples, unsupervised PCA and hierarchical clustering (HCA; Ward's method, Euclidean metric) analysis were performed. Region-specific gene biomarkers were identified based on the loadings that describe the relative contributions of the probe sets' effect on the decomposition of the data. In a second stage, loadings (for probe sets) and scorings (for samples) were mapped onto a biplot to determine region-specific gene regulation. The selected genes were then crossed with the highly variable ones identified in the GSE1152 data set, publicly available at the GEO repository. All analyses were performed using R Software and Bioconductor packages (Gentleman et al. 2004). The statistical significance of differential gene expression was determined using a modified ANOVA according to the Global Error Assessment procedure (Mansourian et al. 2004). Finally, to bypass the problem of multiple Present (P) and Absent (A) calls for the same gene/condition, we computed a virtual chip using the robust average of the probe sets' values of the same gene in the different chips and then computed the P/A calls using the Affymetrix MAS 5.0 algorithm.

Bioinformatics analyses

Probe sets were annotated using ENSEMBL v28 and NetAffx (Affymetrix) (grade 1), preferring the former if discordant (grade 2) or the latter if the former was missing (grade 3). Probe sets representing unknown sequences were tentatively annotated (grade 4) by protein domain recognition analysis using HMMer against Pfam (http://pfam.sanger.ac.uk/) and with PSI-Blast (Altschul et al. 1997). The level of confidence of the annotation was expressed as an E value, i.e., the probability of detecting the similarity by chance, as calculated by PSI-Blast (cutoff = 10E-5). Gene expression profiles were represented using Spotfire Decision Site for Functional Genomics 7.3 (TIBCO Spotfire Inc., Palo Alto, CA). An automated literature search based on WordStat v4.0.19 and SimStat v2.5.1 software (Provalis Research, Montreal, CA), followed by manual refinement, was used to identify genes that had never been described in the GI tract.

Results and discussion

Global analysis of the gastrointestinal tract transcriptome

Total mRNAs prepared from the biopsies of antrum, duodenum, jejunum, ileum, and colon of two males and two females were independently hybridized to the Affymetrix gene chip U133A, which contains 22,283 probe sets representing 14,593 UniGene clusters for a total of 20 microarrays used. Villin (205506_at), occludin (209925_at), and the transcription factor Elf3 (201510_at), three genes expressed exclusively in the epithelium, were scored present and highly expressed in all regions. On the other hand, CD80 (207176_s_at), CD86 (205685_at), and CD14 (201743_at), expressed in immune cells, and epimorphin (207346_at), expressed in mesenchymal cells, were scored absent and/or with very low hybridization intensities. Therefore, we can assume that the expression profiles produced in this study represent mainly epithelial cells. A blind unsupervised approach was used to assess the samples' separability while reducing the amount of noise by complementing PCA and HCA analyses. The representation of the data on the two first principal components clearly showed two possible outliers, one from antrum gathering with ileum samples and one from ileum gathering with jejunum samples (data not shown). These two samples were removed from further analysis since deviation of a single representative point from the cluster of its peers, at a distance as great as ten times the standard deviation of the related cluster, seemed suspicious, and the chance of this representing a real biological effect was estimated to be vanishing. HCA and PCA performed on the remaining 18 samples gave consistent and overlapping results and provided the basis to further restrain the analysis to those probe sets, out of the original 22,283, that varied most significantly among them. A magnitude of every probe set contribution was estimated and the original data set was reduced to 199 significant probe sets (Supplementary Table 1). These probe sets, selected on the basis of their important loadings (high loading corresponds to high variance), are those mainly contributing to the observed intracluster energy and therefore are responsible for sample separation. When carrying out HCA and PCA on these probe sets, we found that samples were not separable according to gender but were rather clearly stratified according to the gastrointestinal region (Fig. 1a, b). PCA was able to capture more than 85% of the information contained in the raw data set. Sample separation was particularly evident for antrum, colon, and ileum, while duodenum and jejunum were somewhat more similar. This was further highlighted by analyzing the region-specific modulation of the 199 selected probe sets through scorings and loadings mapping on a biplot (Fig. 1c). Thirty-eight probe sets were modulated in antrum, 20 in ileum, and 33 in colon. The identification of probe sets selective of duodenum and jejunum turned out to be more difficult. Of a total of 108 probe sets specifically modulated in these two compartments, we could identify 21 probe sets in duodenum and 15 in jejunum, while the remaining 72 were jointly modulated in both tissues. No probe set expressed in all regions but one was found. All of the 199 probe sets were found to be statistically significantly differentially expressed by using the Global Error Assessment (GEA) model ($P < 0.001$), except for 220038_at which was therefore excluded from further examination.

Identification and expression analysis of healthy intestinal regions' selective genes

The 198 probe sets selected above were assigned to 166 unique genes (Fig. 2) that we defined as region-selective but not necessarily region-specific, with some expressed in more than one GI segment according to Affymetrix present calls (Supplementary Table 1). Region-selective genes are preferentially, but not exclusively, expressed in a given region and are likely to be involved in its major physiological functions. Region-specific genes are restricted to that region and may be used to distinguish it (Hsiao et al. 2001). Automated literature mining followed by manual refinement was used for the 166 genes to discriminate between novel and existing information. One hundred sixteen genes were found to be associated with known intestinal functions and their expression profile corresponded to previously reported data. In particular, the expression profile of most of these genes reflected intestinal biological functions that are anatomically distributed, including digestion in the upper regions and absorption and excretion in the middle and lower parts of the GI tract. This validated our gene-selection approach and these genes were not further investigated. Thirty-one genes had already been described in the GI tract but their proximal-to-distal profile had not been documented. These are NMU, MUC1, MUC6, FOXA2, GATA4, ORM1, and CRIP2 selective of antrum; TM4SF4 and SERPINB5 selective of duodenum; GPR172B, SFRP5, HTR1D, BST1, and RASA4 selective of duodenum and jejunum; CLDN15 selective of jejunum; UGT2B17, CFTR, MUC2, ZG16, and PPP1R12A selective of ileum; and PDE9A, HOXA9, HOXA6, PYGB, POSTN, CEACAM5, CD24, TSPAN-1, CNNM4, HK2, and RARRES1 selective of colon. Finally, 19 genes had never been reported in the human GI tract. These include SEMA6A, a semaphorin involved in the control of actin filament dynamics (Klostermann et al. 2000) and possibly in pathogen invasion (Laurent et al. 1999) and found to be ileum-selective; PCDH21, a cadherin suggested to be involved in

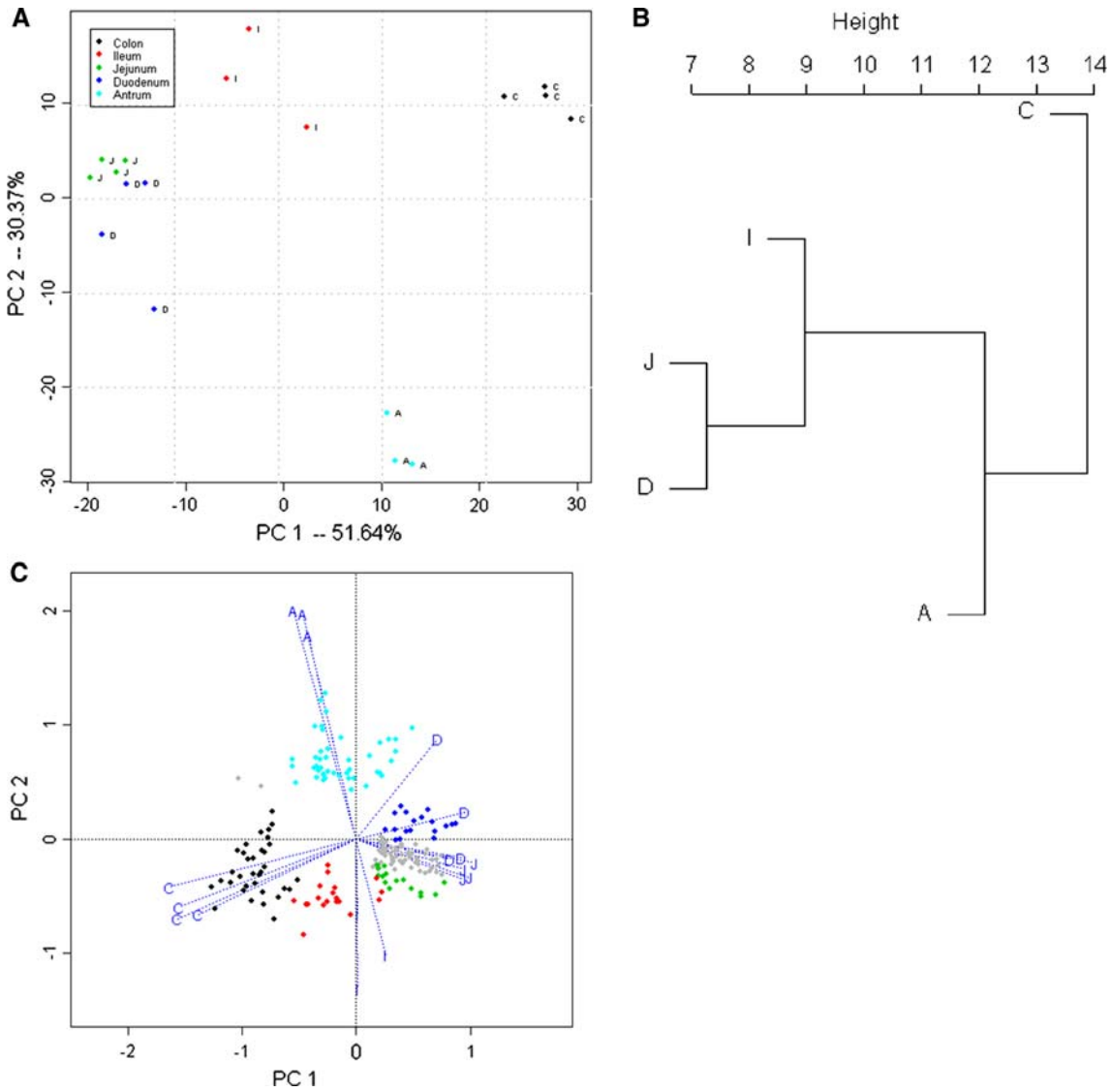

Fig. 1 Clustering of gastrointestinal regions' transcriptomes based on the 199 probe sets responsible for sample separability. Principal components analysis (PCA) (**a**) and hierarchical clustering (Ward's method) (**b**) analysis were conducted on the 18 samples, each individually represented on the graphs. The scale of **b** is the sum of the squares of the euclidean distance (the unit) for the cluster and it is based on a differential matrix. Consistency between A (PCA) and B supported the tree which was generated by hierarchical clustering. The scorings (samples) and loadings (probe sets) from PCA analysis were then mapped on the same plot, called a biplot (**b**). As a consequence of the matrix factorization X (log transformed and centered on 0), the modulation of any probe set can be geometrically inferred from its position on the biplot. Accordingly, when a probe set lies far from the origin and close to one chip, it is upregulated in this chip (gastrointestinal segment sample) compared to all the others and vice versa. The highest the distance of a probe set from the origin, the highest its modulation. The magnitude of modulation for each probe set is given as a numerical value (Norm value, Fig. 2). *A* antrum, *D* duodenum, *J* jejunum, *I* ileum, *C* colon

the formation of neuronal networks in the olfactory bulb (Nakajima et al. 2001) and found to be colon-specific; MS4A12, a member of the membrane-spanning 4A family thought to be involved in signal transduction (Liang and Tedder 2001) which is colon-specific; CLIC3, a nuclear chloride channel found to be specific of antrum; and CLCA4, a Ca^{2+}-activated chloride channel which is colon selective. Six of the 19 genes were unknown genes that were tentatively annotated as explained in Materials and methods. Among these we identified, based on protein sequence analysis, FLJ22880 as a new putative member ($E = 3.8e^{-134}$) of the L6 family of transmembrane proteins (Maecker et al. 1997; Wright et al. 2000) that is selective of small intestine and likely to be implicated in signal transduction events, and a dihydroxyacetone kinase (DAK-1, $E = 7.5e^{-128}$; DAK-2, $E = 4.3e^{-69}$) that catalyzes the first limiting step of glycerol utilization (glycerone to glycerone phosphate, KEGG map00561, http://www.genome.jp/kegg/)

Expression pattern	Regional selectivity	Gene Description	Grade	Gene Symbol	Unigene	Probe set ID	Norm value
A	A	Mucin 5, subtypes A and C, tracheobronchial/gastric	3	MUC5AC	Hs.534332	214385_s_at	2.83
A	A	Gastrokine 1 precursor (18 kDa antrum mucosa protein)	1	GKN1	Hs.69319	220191_at	2.59
A	A	Trefoil factor 2 precursor	1	TFF2	Hs.2979	214476_at	2.44
A	A	Prostate stem cell antigen precursor	1	PSCA	Hs.379010	205319_at	2.30
A	A	Gastricsin precursor	1	PGC	Hs.1867	205261_at	2.10
A	A	Annexin A10 (Annexin 14)	1	ANXA10	Hs.188401	210143_at	2.09
A	A	Gastrin precursor	1	GAS	Hs.2681	208138_at	2.05
A	A	S-100P protein	1	S100P	Hs.2962	204351_at	2.04
A	A	Triacylglycerol lipase, gastric precursor	1	LIPF	Hs.523130	206334_at	1.97
A	A	Mucin 1, transmembrane	3	MUC1	Hs.89603	213693_s_at	1.89
A	A	Claudin-18	1	CLDN18	Hs.240182	214135_at	1.88
A	A	Aldehyde dehydrogenase, dimeric NADP-preferring	1	ALDH3A1	Hs.531682	205623_at	1.85
A	A	Mucin 6, gastric	1	MUC6	Hs.528432	214133_at	1.85
A	A	Pepsin A precursor	1	PGA5	Hs.432854	213265_at	1.81
A	A	**Tumor-associated calcium signal transducer 2 (Cell surface glycoprotein Trop-2)**	1	TACSTD2	Hs.23582	202286_s_at	1.80
A	A	Intrinsic factor precursor (IF) (INF) (Gastric intrinsic factor)	1	GIF	Hs.110014	207033_at	1.76
A	A	Potassium-transporting ATPase beta chain	1	ATP4B	Hs.434202	207546_at	1.67
A	A	Trefoil factor 1 precursor	1	TFF1	Hs.162807	205009_at	1.64
A	A	Beta-microseminoprotein precursor	1	MSMB	Hs.255462	207430_s_at	1.62
A	A	Neuromedin U-25 precursor	1	NMU	Hs.418367	206023_at	1.55
A	A	Lactotransferrin precursor	1	LTF	Hs.529517	202018_s_at	1.55
A	A	G protein-coupled receptor family C group 5 member B (Retinoic acid induced gene 2 protein)	1	GPRC5B	Hs.148685	203632_s_at	1.53
A	A	Potassium-transporting ATPase alpha chain 1 (Gastric H+/K+ ATPase alpha subunit)	1	ATP4A	Hs.36992	207139_at	1.48
A	A	Hepatocyte nuclear factor 3-beta (Forkhead box protein A2)	1	FOXA2	Hs.155651	210103_s_at	1.41
A	A	Cathepsin E precursor	1	CTSE	Hs.1355	205927_s_at	1.41
A	A	**Chloride intracellular channel protein 3**	1	CLIC3	Hs.64746	219529_at	1.41
A	A	Sulfotransferase family, cytosolic, 1C, member 1	3	SULT1C1	Hs.436123	211470_s_at	1.37
A	A	Potassium voltage-gated channel subfamily E member 2	1	KCNE2	Hs.551521	221095_s_at	1.34
A	A	Alpha-1-acid glycoprotein 2 precursor (Orosomucoid 2)	2	ORM1	Hs.494894	205040_at	1.33
A	A	Tescalcin	1	TSC	Hs.525709	218872_at	1.32
A	A	**Dual specificity protein phosphatase 4**	1	DUSP4	Hs.417962	204015_s_at	1.30
A	A	**No description**	4	—	Hs.49117	214046_at	1.30
A	A	Transcription factor GATA-4	1	GATA4	Hs.243987	205517_at	1.27
A	A	Somatostatin precursor	2	SST	Hs.12409	213921_at	1.23
A	A	Clusterin precursor	1	CLU	Hs.436657	208791_at	1.11
A	A	Cysteine-rich protein 2	1	CRIP2	Hs.534309	208978_at	0.99

Fig. 2 Proximal-to-distal expression pattern of 166 region-selective genes along the GI tract. The 199 probe sets responsible for sample clustering (Fig. 1) were found to correspond to 166 genes selective of **a** antrum (36 genes); **b** duodenum (14 genes), duodenum and jejunum (57 genes), and jejunum (14 genes); **c** ileum (16 genes) and colon (29 genes). Their expression pattern along the gastrointestinal tract is visualized in the heat map that was constructed using the ln-transformed average hybridization intensities calculated from three to four replicates (individuals). The gene description corresponds to the ENSEMBL and Affymetrix designation (grade 1), ENSEMBL only (grade 2), and Affymetrix only (grade 3), or was tentatively assigned as explained in the text (grade 4). The normalization (Norm) value, calculated by PCA, represents the importance of each gene in distinguishing the related gut segment from the others; values close to or higher than 2 correspond to high tissue selectivity. Bold indicates genes not previously described to be expressed in the corresponding gut region based on automated literature search, as explained in Materials and methods. *A* antrum, *D* duodenum, *J* jejunum, *I* ileum, *C* colon, *na* not applicable

and that has not yet been characterized in human and has been found to be selective of proximal small intestine. A putatively novel transcription factor containing JmjC domains (FLJ13798, $E = 7.8e^{-05}$) was identified as selective of proximal small intestinal regions but has not been further investigated since it was called absent by the Affymetrix algorithm. The colon-specific probe set 220724_at was assigned to a nuclease/phosphatase which, despite the low level of confidence ($E = 0.16$), was confirmed by fold recognition analysis (not shown). Finally, two probe sets (214046_at and 220645_at) remained unassigned, although the gene represented by 220645_at is called 99% similar to a brush border protein gene belonging to the same UniGene cluster. Based on UniGene's EST Profile Viewer, these 19 genes indeed were confirmed to be expressed in the GI segment to which they were found to be associated, except for CLIC3, FLJ13798, CML2, and SEMA6A for which complete information was not available.

Subsequent particular analysis of gene expression modulation showed that the antrum was characterized by the modulation of genes encoding for several gastric proteins such as trefoil factors, progastricsin, and mucin 5AC (Fig. 2a), as well as a gene encoding for the cytosolic sulfotransferase SULT1C1, a finding supported by previous research reporting sulfotransferase activity in the antral and body mucosa of human and rat stomachs. This enzyme catalyzes the transfer of a sulfate ester group from 3′-phosphoadenosine 5′-phosphosulfate to the carbohydrate chain of gastric mucus glycoproteins (Carter et al. 1988; Liau et al. 1983). Another important finding was the identification of the β-microseminoprotein isoform-a precursor. This is a small disulfide-rich protein with unknown function that is present in the secretions of most mucosal surfaces in the body, including the stomach where it is secreted in the antrum by endocrine cells (Weiber et al. 1997). The protein has been investigated as a biomarker of gastric carcinoids (Weiber et al. 1999).

B

Expression pattern A D J I C	Regional selectivity	Gene Description	Grade	Gene Symbol	Unigene	Probe set ID	Norm value
	D	Interleukin-1 receptor, type II precursor	1	IL1R2	Hs.25333	205403_at	2.55
	D	Vitamin D-dependent calcium-binding protein, intestinal (CABP) (Calbindin D9K)	1	S100G	Hs.639	207885_at	1.95
	D	Glutamate carboxypeptidase II	1	FOLH1	Hs.380325	205860_x_at	1.87
	D	Enteropeptidase precursor	1	PRSS7	Hs.149473	207638_at	1.54
	D	Transmembrane 4 superfamily member 4 (Intestine and liver tetraspan membrane protein)	1	TM4SF4	Hs.133527	209937_at	1.52
	D	Gastric inhibitory polypeptide precursor (GIP) (Glucose-dependent insulinotropic polypeptide)	1	GIP	Hs.1454	207899_at	1.32
	D	Adenosine deaminase	1	ADA	Hs.407135	204639_at	1.16
	D	Procholecystokinin precursor	1	CCK	Hs.458426	205827_at	1.07
	D	Arylacetamide deacetylase	1	AADAC	Hs.506908	205969_at	1.06
	D	Nuclear receptor 0B2 (Orphan nuclear receptor SHP)	1	NR0B2	Hs.427055	206410_at	1.06
	D	Lactase-phlorizin hydrolase precursor	1	LCT	Hs.551506	206945_at	1.00
	D	NPC1 (Niemann-Pick disease, type C1, gene)-like 1	1	NPC1L1	Hs.512147	220106_at	0.88
	D	Serine (or cysteine) proteinase inhibitor, clade B (ovalbumin), member 5	3	SERPINB5	Hs.55279	204855_at	0.76
	D	Cytochrome P450 2C9	1	CYP2C9	Hs.282624	220017_x_at	0.61
	JD	Cytochrome b reductase 1; duodenal cytochrome b	1	CYBRD1	Hs.221941	217889_s_at	1.80
	JD	Neprilysin	1	MME	Hs.307734	203434_s_at	1.69
	JD	Apolipoprotein A-IV precursor	1	APOA4	Hs.1247	206894_at	1.64
	JD	Apolipoprotein A-I	3	APOA1	Hs.534983	217073_x_at	1.58
	JD	Deoxyribonuclease I precursor	1	DNASE1	Hs.549030	210165_at	1.57
	JD	Oligopeptide transporter, (Intestinal H+/peptide cotransporter)	1	SLC15A1	Hs.436893	207254_at	1.40
	JD	Pantetheinase precursor (Vanin 1) (Tiff66)	1	VNN1	Hs.12114	205844_at	1.38
	JD	Apolipoprotein C-III precursor (Apo-CIII)	1	APOC3	Hs.534984	205820_s_at	1.34
	JD	Aldolase B, fructose-bisphosphate	3	ALDOB	Hs.530274	217238_s_at	1.25
	JD	**Complement component C8 gamma chain precursor**	1	C8G	Hs.1285	210324_at	1.24
	JD	Microsomal triglyceride transfer protein large subunit precursor	1	MTP	Hs.195799	205675_at	1.24
	JD	**L6 membrane protein**	4	FLJ22800	Hs.156652	220639_at	1.21
	JD	Trehalase precursor	1	TREH	Hs.129712	207378_at	1.17
	JD	Amine oxidase [flavin-containing] B (EC 1.4.3.4) (Monoamine oxidase) (MAO-B)	1	MAOB	Hs.46732	204041_at	1.15
	JD	Apolipoprotein C-II precursor (Apo-CII)	1	APOC2	Hs.75615	204561_x_at	1.09
	JD	ATP-binding cassette, sub-family G, member 5 (Sterolin-1)	1	ABCG5	Hs.132992	220383_at	1.07
	JD	G protein-coupled receptor 172B; putative G-protein coupled receptor GPCR42	1	GPR172B	Hs.110128	220756_s_at	1.05
	JD	Pancreatic lipase related protein 2 precursor	1	PNLIPRP2	Hs.423598	211766_s_at	1.02
	JD	Defensin 6 precursor (Defensin, alpha 6)	1	DEFA6	Hs.711	207814_at	0.97
	JD	Cytochrome P450 3A4	1	CYP3A4	Hs.442527	205999_x_at	0.94
	JD	Solute carrier family 13, member 2 (Na(+)/dicarboxylate cotransporter 1)	1	SLC13A2	Hs.102307	207567_at	0.93
	JD	Solute carrier family 2, facilitated glucose transporter, member 5 (Fructose transporter)	1	SLC2A5	Hs.530003	204430_s_at	0.93
	JD	Glutathione S-transferase A1	1	GSTA1	Hs.446309	203924_at	0.90
	JD	Lithostathine 1 alpha precursor (PSP) (REG)	1	REG1A	Hs.49407	209752_at	0.86
	JD	Chemokine (C-C motif) ligand 25	3	CCL25	Hs.310511	206988_at	0.86
	JD	Glycerol-3-phosphate dehydrogenase 1 (soluble)	3	GPD1	Hs.524418	213706_at	0.85
	JD	Sodium/glucose cotransporter 1	1	SLC5A1	Hs.1964	206628_at	0.81
	JD	Canalicular multispecific organic anion transporter 1(Multidrug resistance-associated protein 2)	1	ABCC2	Hs.368243	206155_at	0.81
	JD	Secreted frizzled-related protein 5; secreted apoptosis related protein 3	1	SFRP5	Hs.279565	207468_s_at	0.80
	JD	B(0,+)-type amino acid transporter 1 (B(0,+)AT)	1	SLC7A9	Hs.408567	220135_s_at	0.80
	JD	Carbamoyl-phosphate synthetase I	1	CPS1	Hs.149252	204920_at	0.78
	JD	Defensin 5 precursor (Defensin, alpha 5)	1	DEFA5	Hs.72887	207529_at	0.77
	JD	Pancreatitis-associated protein 1 precursor	1	PAP	Hs.423	205815_at	0.75
	JD	**TF-like jumonji**	4	FLJ13798	Hs.145717	220070_at	0.69
	JD	Sodium- and chloride-dependent transporter XTRP3 (Neurotransmitter transporter rB21A homolog)	1	SLC6A20	Hs.413095	219614_s_at	0.68
	JD	Ketohexokinase (Hepatic fructokinase)	1	KHK	Hs.159525	211028_s_at	0.66
	JD	Glucose-6-phosphatase	1	G6PC	Hs.212293	206952_at	0.65
	JD	Solute carrier family 25 (mitochondrial carrier; ornithine transporter) member 15	3	SLC25A15	Hs.442781	218653_at	0.64
	JD	Solute carrier family 4, sodium bicarbonate cotransporter, member 7	1	SLC4A7	Hs.250072	209884_s_at	0.63
	JD	Deleted in malignant brain tumors 1 isoform b precursor	1	DMBT1	Hs.279611	208250_s_at	0.62
	JD	Sodium/nucleoside cotransporter 2	1	SLC28A2	Hs.367833	207249_s_at	0.62
	JD	**Putative N-acetyltransferase Camello 2**	3	CML2	Hs.458287	206964_at	0.61
	JD	Xaa-Pro aminopeptidase 2 precursor	2	XPNPEP2	Hs.170499	216910_at	0.61
	JD	Monocarboxylate transporter 5 (MCT 5) (MCT 4)	1	SLC16A4	Hs.351306	205234_at	0.60
	JD	Cytochrome P450 3A7	1	CYP3A7	Hs.111944	205939_at	0.59
	JD	Galactosylceramide sulfotransferase	1	GAL3ST1	Hs.17958	205670_at	0.59
	JD	Ornithine carbamoyltransferase, mitochondrial precursor	1	OTC	Hs.117050	207200_at	0.58
	JD	Cytochrome P450 27	1	CYP27A1	Hs.516700	203979_at	0.57
	JD	**Transcription factor Maf (Proto-oncogene c-maf)**	1	MAF	Hs.134859	209347_s_at	0.57
	JD	5-hydroxytryptamine 1D receptor (5-HT-1D) (Serotonin receptor 1D)	1	HTR1D	Hs.121482	207368_at	0.56
	JD	Platelet glycoprotein IV (CD36 antigen)	1	CD36	Hs.120949	209555_s_at	0.56
	JD	**Dak 1 and 3 DOMAINS-containing**	4	DKFZP586B1621	Hs.6278	218688_at	0.56
	JD	ADP-ribosyl cyclase 2 precursor	1	BST1	Hs.169998	205715_at	0.53
	JD	**Hepatocyte growth factor-like protein precursor**	1	MST1	Hs.349110	205614_x_at	0.52
	JD	Secretin precursor	1	SCT	Hs.553503	221198_at	0.51
	JD	**Cell death activator CIDE-B (Cell death-inducing DFFA-like effector B)**	1	CIDEB	Hs.448590	221188_s_at	0.49
	JD	Ras GTPase-activating protein 4	1	RASA4	Hs.520599	208534_s_at	0.48
	J	Alkaline phosphatase, intestinal precursor	1	ALPI	Hs.37009	207140_at	1.89
	J	Solute carrier family 2, facilitated glucose transporter, member 2	1	SLC2A2	Hs.167584	206535_at	1.72
	J	maltase-glucoamylase (alpha-glucosidase)	3	MGAM	Hs.122785	206522_at	1.67
	J	Dipeptidyl peptidase IV	1	DPP4	Hs.368912	203716_s_at	1.46
	J	Apolipoprotein B-100 precursor	1	APOB	Hs.120759	205108_s_at	1.19
	J	Angiotensin I converting enzyme 2 precursor; angiotensin converting enzyme-like protein	1	ACE2	Hs.178098	219962_at	1.18
	J	Sucrase-isomaltase, intestinal	1	SI	Hs.429596	206664_at	1.13
	J	Glutamyl aminopeptidase	1	ENPEP	Hs.435765	204844_at	0.98
	J	Annexin A13 (Annexin XIII) (Annexin, intestine-specific) (ISA)	1	ANXA13	Hs.181107	208323_s_at	0.94
	J	Meprin A beta-subunit precursor	1	MEP1B	Hs.194777	207251_at	0.87
	J	Claudin-15	1	CLDN15	Hs.38738	219640_at	0.84
	J	Cytochrome P450, family 4, subfamily F, polypeptide 2	3	CYP4F2	Hs.101	210452_x_at	0.78
	J	PDZ domain containing 1	1	PDZK1	Hs.143293	205380_at	0.73

Fig. 2 continued

Expression pattern A D J I C	Regional selectivity	Gene Description	Grade	Gene Symbol	Unigene	Probe set ID	Norm value
	I	Calcineurin B homologous protein 2	1	LOC63928	Hs.178589	206149_at	1.47
	I	Multidrug resistance protein 3 (P-glycoprotein 3)	1	ABCB1 /// ABCB4	Hs.489033	209994_s_at	1.43
	I	**Semaphorin 6A**	3	SEMA6A	Hs.156967	215028_at	1.34
	I	UDP-glucuronosyltransferase 2B17 precursor	1	UGT2B17	Hs.183596	207245_at	1.34
	I	Mucin and cadherin-like	3	MUCDHL	Hs.165619	220075_s_at	1.29
	I	Calcium activated chloride channel 1 precursor	1	CLCA1	Hs.194659	210107_at	1.28
	I	Fatty acid-binding protein, liver (L-FABP)	1	FABP1	Hs.380135	205892_s_at	1.16
	I	Chloride anion exchanger (DRA protein) solute carrier family 26, member 3	1	SLC26A3	Hs.1650	206143_at	1.11
	I	Cystic fibrosis transmembrane conductance regulator (CFTR)	1	CFTR	Hs.489786	205043_at	1.09
	I	Gastrotropin (GT) (Ileal lipid-binding protein) (ILBP)	1	FABP6	Hs.519719	210445_at	2.03
	I	Peptide YY	3	PYY	Hs.169249	207080_s_at	1.71
	I	Mucin 2, intestinal/tracheal	3	MUC2	Hs.315	204673_at	1.55
	I	Zymogen granule protein 16	1	ZG16	Hs.184507	214142_at	1.52
	I	ADAM-like, decysin 1; disintegrin protease; decysin	1	ADAMDEC1	Hs.521459	206134_at	1.32
	I	TRAF and TNF receptor-associated protein	1	TTRAP	Hs.403010	202266_at	1.11
	I	Protein phosphatase 1 regulatory subunit 12A	1	PPP1R12A	Hs.49582	201604_s_at	0.87
	C	**Calcium activated chloride channel 4**	1	CLCA4	Hs.546343	220026_at	3.01
	C	Carcinoembryonic antigen-related cell adhesion molecule 7	3	CEACAM7	Hs.74466	211848_s_at	2.91
	C	Aquaporin 8	1	AQP8	Hs.176658	206784_at	2.52
	C	NADPH oxidase homolog 1 (NOX-1)	1	NOX1	Hs.132370	206418_at	2.38
	C	**DNA-binding protein SATB2**	1	SATB2	Hs.516617	213435_at	2.34
	C	**Membrane-spanning 4-domains subfamily A member 12**	1	MS4A12	Hs.272789	220834_at	2.32
	C	Carbonic anhydrase XII	3	CA12	Hs.210995	204508_s_at	2.30
	C	Carbonic anhydrase IV precursor	1	CA4	Hs.89485	206208_at	2.28
	C	Guanylin precursor (Guanylate cyclase activator 2A)	1	GUCA2A	Hs.778	207003_at	2.18
	C	Hexokinase, type II	1	HK2	Hs.406266	202934_at	2.18
	C	Glucagon precursor	1	GCG	Hs.516494	206422_at	2.17
	C	Mucin 5B precursor	1	MUC5B	Hs.523395	213432_at	2.12
	C	**Nuclease**	4	FLJ21511	Hs.479703	220724_at	2.11
	C	**Protocadherin 21 precursor**	1	PCDH21	Hs.137556	213369_at	2.07
	C	Carbonic anhydrase I	1	CA1	Hs.23118	205950_s_at	2.07
	C	Mucin 4 isoform d	1	MUC4	Hs.369646	204895_x_at	2.00
	C	Phosphodiesterase 9A	3	PDE9A	Hs.473927	205593_s_at	1.98
	C	Homeobox protein Hox-A9 (Hox-1G)	1	HOXA9	Hs.127428	214651_s_at	1.87
	C	Homeobox protein Hox-B6	1	HOXB6	Hs.98428	205366_s_at	1.85
	C	Glycogen phosphorylase, brain form	1	PYGB	Hs.368157	201481_s_at	1.83
	C	**no description**	na	C11orf33	Hs.179100	220645_at	1.82
	C	**Coiled-coil domain containing protein 2 (Capillary morphogenesis protein-1)**	1	CCDC2	Hs.145402	219174_at	1.77
	C	Periostin precursor Osteoblast specific factor 2	1	POSTN	Hs.136348	210809_s_at	1.69
	C	Carcinoembryonic antigen-related cell adhesion molecule 5	3	CEACAM5	Hs.220529	201884_at	1.69
	C	Signal transducer CD24 precursor	1	CD24	Hs.375108	209772_s_at	1.66
	C	Tetraspanin 1 (Tspan-1) (Tetraspan NET-1) (Tetraspanin TM4-C)	1	TSPAN-1	Hs.38972	209114_at	1.66
	C	Cyclin M4; ancient conserved domain protein 4	1	CNNM4	Hs.175043	218900_at	1.66
	C	Sulfate transporter	1	SLC26A2	Hs.302738	205097_at	1.65
	C	Retinoic acid receptor responder protein 1 (Tazarotene-induced gene 1 protein)	1	RARRES1	Hs.131269	206392_s_at	1.62

Fig. 2 continued

The whole experimental approach was clearly validated by finding that the unique probe set on the HG-U133A Affymetrix chip that was annotated to be antrum-specific was retrieved as being specifically modulated in this same region (probe set ID: 220191-at). The duodenum and jejunum (Fig. 2b) were characterized by the modulation of a large spectrum of proteolytic enzymes, a number of lipoprotein precursors, and several solute carrier proteins (transporters). A number of detoxifying enzymes that belong to the cytochrome P450 family were also modulated in these regions. In what concerned duodenum tissue only, modulation of the NPC1 gene was found remarkable. A deficiency in the function of this gene is related to the Niemann-Pick disease type C that is due to defective cholesterol intracellular transport (Liscum and Klansek 1998). The jejunum was characterized by peptido- and glycohydrolytic activities due to the specific modulation of the related enzymes such as dipeptidylpeptidase, aminopeptidase, glucoamylase, and isomaltase. Several interesting genes appeared to be modulated in both duodenum and jejunum, such as vanin-1, which is involved in the biosynthesis of vitamin B (Martin et al. 2004), and sterolin-1, which prevents dietary noncholesterol sterols from being absorbed (Klett and Patel 2003). Among genes selective of ileum (Fig. 2c), we found MUC2, which codes for a mucin associated with high colonic expression (Tytgat et al. 1994). Analysis of the corresponding probe set position on the biplot (Fig. 1c) revealed that it is almost equidistant from ileum and colon, highlighting similar expression in the two regions. Accordingly, mucin 2, besides being dominant in colon mucins, could also be one of the main ileal mucins. An important novelty came from scrutiny of other genes selective of ileum, as we found evidence for the important modulation of semaphorin 6A1, a protein involved in neuronal development and regeneration during apoptosis that is discussed further below. The colon (Fig. 2c) was characterized by an elevated modulation of genes encoding for carbonic anhydrase precursors as well as for carcinoembryonic antigen-related cell adhesion molecules. Few other genes appeared to be modulated, e.g., aquaporin and the osteoblast specific factor 2.

Our analysis was validated once more by the finding of the known differential expression of the claudin family of

tight junction proteins involved in paracellular transport (Rahner et al. 2001). For instance, the expression of claudin 4 was found to be highly restricted to the colon, whereas claudin 18 was mainly expressed in the antrum. Knowledge of these expression patterns is significant as it has been proposed that they underlie differences in the paracellular permeability properties (Niimi et al. 2001).

Among novel genes and genes not previously known to be expressed in the GI tract, we found several transmembrane proteins that seem to be important to gut regions in a specific manner. These include SEMA6A, FLJ22800, TSPAN1, and MS4A12. This has possible implications in the recognition of non-self molecules such as nutrients, pathogenic and commensal bacteria, or any other antigen, and subsequently generated signaling events in the enterocyte. For example, it has been suggested that SEMA6A plays a role in retrograde signaling to cytoskeletal elements by regulating proteins like Ena/VASP (Klostermann et al. 2000). Interestingly, these proteins mediate intracellular movement of *L. monocytogenes* by linking the bacterium to the actin tail (Laurent et al. 1999); therefore, semaphorin 6A may play a role in pathogen infection in the small intestine. We suggest that FLJ22800 is a new L6 type of protein, based on protein sequence analysis. L6 proteins are surface proteins characterized by four transmembrane domains and constitute a family of four members: L6, IL-TMP, TM4SF5, and L6D (Wright et al. 2000). The biological function of these proteins is not completely understood, although they have been implicated in signal transduction events. Moreover, it has been suggested that they mediate tumor cell metastasis and cell adhesion (Kao et al. 2003). All except L6D are represented on the Affymetrix U133A chip used in this study so we could therefore examine their expression pattern. They all appear to have very low expression in colon and enriched expression in the proximal small intestine. It is notable that IL-TMP has been identified in this study as a descriptor of duodenum. TM4SF5 is the only one expressed in the antrum, while L6 seems to be the only one not expressed in the GI tract, even though this should be confirmed by independent assays. L6 proteins were previously considered to be tetraspanins (Maecker et al. 1997; Marken et al. 1992) because of their similar topological features. Tetraspanins are transmembrane proteins that support and stabilize the formation of signaling complexes (Maecker et al. 1997). The expression of several tetraspanins as well has been shown to be highly regulated along the GI tract (Okochi et al. 1999). We found TSPAN1 (transmembrane protein tetraspanin-1) to be a gene selective of colon. Of all members of the tetraspanin family, TSPAN-1 is most similar to CD81 (Todd et al. 1998), which is expressed in olfactory neurons. The specific role of TSPAN-1 in colon is still unknown. Finally, MS4A12 is a member of the membrane-spanning 4A gene family, whose expression pattern and biological role are unknown even though it has been suggested to be involved in signal transduction, as are the other family members (Liang and Tedder 2001). The high expression level and modulation value in colon, as well as the fact that no ESTs have been found in any tissue but colon, suggest that this gene is colon-specific.

Taken as a whole, these data represent the first mapping of the healthy adult human gastrointestinal transcriptome.

Identification and expression pattern analysis of ileum and colon selective genes in IBD

To show the effective use of our approach when studying gene expression deviation in gut pathologies, we reanalyzed the GEO data set GSE1152 that contains gene expression profiles corresponding to the ileum and colon of Crohn's, ulcerative colitis (UC), and control patients (Langmann et al. 2004). In that study, the same microarray platform used here was employed. First we explored regional clustering in the IBD condition by applying the approach described above. We carried out a global analysis of the data using the 22,283 probe sets and all the samples crossing the regions (ileum and colon) and conditions (healthy, Crohn's, and UC). We reduced the original data set to only the significant probe sets (329) using their magnitudes. When performing PCA on this subset we were able to see a clear stratification according to the disease rather than to the region type (Fig. 3), which confirmed previous observations (Langmann et al. 2004). This withstanding, the gene expression profile in ileum from UC patients tends to approach that of the control samples, which shows that the transcriptome of this organ is only slightly affected by this disease compared to transverse colon. Finally, 37 of the 45 genes found to be selective of the healthy ileum and colon in our data set (Fig. 2) were modulated in the IBD state, with an absolute fold higher than 1.5 (Supplementary Table 2). Next, we carried out a more detailed analysis by region and by disease. Because we lacked replicates from the GSE1152 data set since samples had been pooled per condition, we were not able to select probe sets statistically; therefore, we took into account the expression level and the density of the probe sets and selected the highly modulated ones, both by intestinal region and by disease, using a binning and skimming technique. We constructed a plot of the $\log_2$fold (disease/healthy) versus the mean $\log_2$expression, thus identifying highly variable probe sets independent from the level of expression. These probe sets were then crossed with the 199 descriptor probe sets identified by our previous analysis (Fig. 1, Supplementary Table 1). The genes thus selected are represented in Fig. 4.

The "by region" analysis (Fig. 4, spots) shows that both Crohn's and ulcerative colitis modulate gastrokine 1

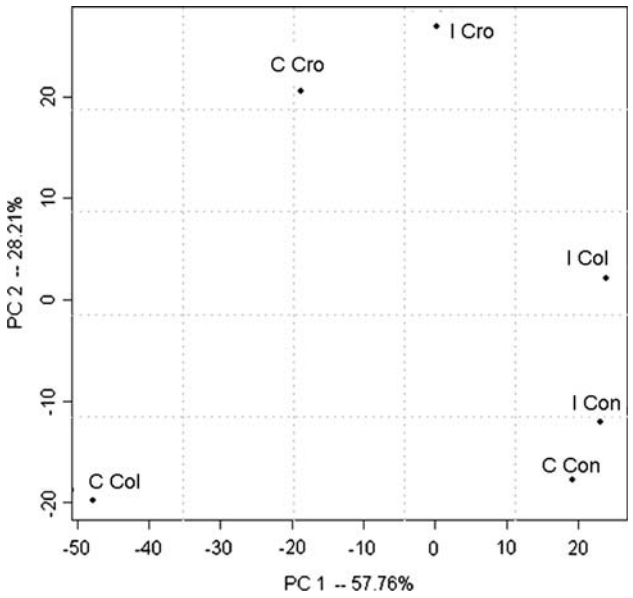

Fig. 3 Clustering of healthy and inflamed ileum and colon transcriptomes. PCA was conducted on the Affymetrix expression data obtained from the healthy and affected ileum and transverse colon of Crohn's and ulcerative colitis patients (data from the GEO data set GSE1552) using the 329 most significant genes (i.e., the genes preselected by PCA to be responsible for samples separability). *I* ileum, *C* colon, *Con* control, *Cro* Crohn's, *Col* colitis

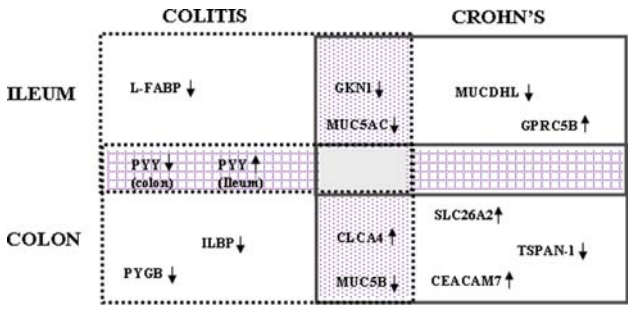

Fig. 4 Genes associated with ulcerative colitis (UC) and/or Crohn's disease in the ileum and/or colon. Genes were identified by a "by region" (ileum and colon) and a 'by disease" (UC and Crohn's) analysis, as explained in the text. *Solid* and *dashed boxes* indicate genes associated with UC and Crohn's, respectively. Genes in the *spots area* were modulated in one of the two regions in both diseases, and genes in the *dotted grid area* were modulated in both regions in one of the diseases. *Arrows* indicate up- and downregulation

(GKN1) and mucin5AC (MUC5AC) in ileum and calcium-activated chloride channel 4 (CLCA4) and mucin5B (MUC5B) in colon. The "by disease" analysis (Fig. 4, dotted grid) shows that none of the region-specific biomarker genes is modulated by Crohn's disease in both ileum and colon, while peptide YY (PYY) is jointly modulated by colitis in both regions. Specifically modulated genes include mucin and cadherin like (MUCDHL) and G protein-coupled receptor family C group 5 member B (GPRC5B) downregulated in Crohn's ileum, liver fatty acid binding protein L-FABP downregulated in colitis ileum, sulfate transporter (SLC26A2) and carcinoembryonic antigen-related cell adhesion molecule 7 (CEACAM7) upregulated in Crohn's colon, tetraspanin-1 (TSPAN-1) downregulated in Crohn's colon, and finally glycogene phosphorylase brain form (PYGB) and gastrotropin or ileal fatty acid binding protein (ILBP) downregulated in colitis colon. Interaction network analysis suggested that these genes are more likely to be involved in independent biological pathways (data not shown). CEACAM7 is a member of the carcinoembryonic antigen (CEA) family and is found only at the apical surface of highly differentiated epithelial cells in the colorectal mucosa (Scholzel et al. 2000). Interestingly, its expression correlates with apoptosis in the normal colon. Increased gene expression in inflamed compared to healthy epithelium may therefore relate to an alteration of the apoptotic pathway accompanied by increased apoptosis, which has recently been suggested to be implicated in the IBD pathology (Verstege et al. 2006). Moreover, The Ca^{2+}-activated chloride channel 4 (CLCA4), which we had found to be the main discriminator of the healthy colon (highest Norm value, Fig. 2), is supposed to upregulate the transmembrane potential and mediate secretion of electrolytes and water (Eggermont 2004). Its upregulation in Crohn's and UC patients may be related to the onset of diarrhea resulting from increased Cl^- secretion. Finally, PYGB, a key enzyme in glycogen degradation involved in calcium and insulin signaling, has been identified in colon. PYGP is not detected in any other region examined in this study and is a strong colon classifier (Norm value = 1.83, Fig. 2). This gene has been proposed to be a marker of carcinogenesis (Tashima et al. 2000). Since IBD patients are at higher risk of developing colorectal cancer, the identification of this gene as a descriptor of UC in colon may be diagnostically relevant.

In summary, PCA could differentiate between the transcriptomes from Crohn's patients and those from colitis and healthy patients based on the selected 329 genes. This withstanding, when looking for descriptors of each region-disease association, we found both genes specific to each of the conditions considered and genes commonly modulated in both diseases and regions (except for ileum and colon of Crohn's patients).

Conclusions

This study provides the first map of gene expression in various compartments of the healthy human adult gastrointestinal tract and a biomathematical information compression approach for the selection of longitudinally differentially expressed genes. The analysis uncovered clear differences in the levels of gene expression among the five

regions under scrutiny in the healthy condition. In particular, it allowed differentiation between the transcriptomes of duodenum and jejunum, which had not been possible in a previous study of gene expression in the murine healthy GI tract (Bates et al. 2002). Furthermore, the approach was applied to an existent data set of gene expression in the intestine of patients affected by Crohn's disease and ulcerative colitis, two manifestations of inflammatory bowel disease. We were therefore able to identify interesting new leads in both the healthy and the disease state. These data furnish a basis for understanding the complex role played by the intestinal epithelium in the maintenance of intestinal homeostasis.

Acknowledgments We thank Maribelle Herranz-Garcia for technical support during tissue sampling and Nadine Porta, Muriel Fiaux, Dominique Donnicola, and Frederic Raymond for technical support with the microarray analysis. We are also thankful to Robert Mansourian for valuable contributions to the statistics and to Ziding Zhang for fold recognition analysis. We are grateful to Anne Donnet, Magali Faure, Clara Garcia-Rodenas, Athula Herath, and Karine Vidal for constructive scientific discussion.

References

Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z et al (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25:3389–3402

Anderle P, Sengstag T, Mutch DM, Rumbo M, Praz V et al (2005) Changes in the transcriptional profile of transporters in the intestine along the anterior-posterior and crypt-villus axes. BMC Genomics 6:69

Andoh A, Saotome T, Sato H, Tsujikawa T, Araki Y et al (2001) Epithelial expression of caveolin-2, but not caveolin-1, is enhanced in the inflamed mucosa of patients with ulcerative colitis. Inflamm Bowel Dis 7:210–214

Backhed F, Ley RE, Sonnenburg JL, Peterson DA, Gordon JI (2005) Host-bacterial mutualism in the human intestine. Science 307:1915–1920

Bates MD, Erwin CR, Sanford LP, Wiginton D, Bezerra JA et al (2002) Novel genes and functional relationships in the adult mouse gastrointestinal tract identified by microarray analysis. Gastroenterology 122:1467–1482

Buisine MP, Desreumaux P, Debailleul V, Gambiez L, Geboes K et al (1999) Abnormalities in mucin gene expression in Crohn's disease. Inflamm Bowel Dis 5:24–32

Buisine MP, Desreumaux P, Leteurtre E, Copin MC, Colombel JF et al (2001) Mucin gene expression in intestinal epithelial cells in Crohn's disease. Gut 49:544–551

Carter SR, Slomiany A, Gwozdzinski K, Liau YH, Slomiany BL (1988) Enzymatic sulfation of mucus glycoprotein in gastric mucosa. Effect of ethanol. J Biol Chem 263:11977–11984

Chou CY, Liu LY, Chen CY, Tsai CH, Hwa HL et al (2008) Gene expression variation increase in trisomy 21 tissues. Mamm Genome 19:398–405

Clavel T, Haller D (2007) Molecular interactions between bacteria, the epithelium, and the mucosal immune system in the intestinal tract: implications for chronic inflammation. Curr Issues Intest Microbiol 8:25–43

Csillag C, Nielsen OH, Borup R, Nielsen FC, Olsen J (2007) Clinical phenotype and gene expression profile in Crohn's disease. Am J Physiol Gastrointest Liver Physiol 292:G298–G304

Diosdado B, Wapenaar MC, Franke L, Duran KJ, Goerres MJ et al (2004) A microarray screen for novel candidate genes in coeliac disease pathogenesis. Gut 53:944–951

Eggermont J (2004) Calcium-activated chloride channels: (un)known, (un)loved? Proc Am Thorac Soc 1:22–27

Fleet JC (2007) Using genomics to understand intestinal biology. J Physiol Biochem 63:83–96

Gentleman RC, Carey VJ, Bates DM, Bolstad B, Dettling M et al (2004) Bioconductor: open software development for computational biology and bioinformatics. Genome Biol 5:R80

Glebov OK, Rodriguez LM, Nakahara K, Jenkins J, Cliatt J et al (2003) Distinguishing right from left colon by the pattern of gene expression. Cancer Epidemiol Biomarkers Prev 12:755–762

Hsiao LL, Dangond F, Yoshida T, Hong R, Jensen RV et al (2001) A compendium of gene expression in normal human tissues. Physiol Genomics 7:97–104

Juuti-Uusitalo K, Maki M, Kainulainen H, Isola J, Kaukinen K (2007) Gluten affects epithelial differentiation-associated genes in small intestinal mucosa of coeliac patients. Clin Exp Immunol 150:294–305

Kao YR, Shih JY, Wen WC, Ko YP, Chen BM et al (2003) Tumor-associated antigen L6 and the invasion of human lung cancer cells. Clin Cancer Res 9:2807–2816

Klett EL, Patel S (2003) Genetic defenses against noncholesterol sterols. Curr Opin Lipidol 14:341–345

Klostermann A, Lutz B, Gertler F, Behl C (2000) The orthologous human and murine semaphorin 6A-1 proteins (SEMA6A-1/Sema6A-1) bind to the enabled/vasodilator-stimulated phosphoprotein-like protein (EVL) via a novel carboxyl-terminal zyxin-like domain. J Biol Chem 275:39647–39653

Langmann T, Moehle C, Mauerer R, Scharl M, Liebisch G et al (2004) Loss of detoxification in inflammatory bowel disease: dysregulation of pregnane X receptor target genes. Gastroenterology 127:26–40

Lapointe LC, Dunne R, Brown GS, Worthley DL, Molloy PL et al (2008) Map of differential transcript expression in the normal large intestine. Physiol Genomics 33:50–64

Laurent V, Loisel TP, Harbeck B, Wehman A, Grobe L et al (1999) Role of proteins of the Ena/VASP family in actin-based motility of *Listeria monocytogenes*. J Cell Biol 144:1245–1258

Liang Y, Tedder TF (2001) Identification of a CD20-, FcepsilonRIb-eta-, and HTm4-related gene family: sixteen new MS4A family members expressed in human and mouse. Genomics 72:119–127

Liau YH, Slomiany BL, Palmer D, Braun DR, Slomiany A et al (1983) Enzymatic sulfation of glycolipids by human gastric mucosa. Digestion 28:132–137

Liscum L, Klansek JJ (1998) Niemann-Pick disease type C. Curr Opin Lipidol 9:131–135

Maecker HT, Todd SC, Levy S (1997) The tetraspanin superfamily: molecular facilitators. FASEB J 11:428–442

Mansourian R, Mutch DM, Antille N, Aubert J, Fogel P et al (2004) The Global Error Assessment (GEA) model for the selection of differentially expressed genes in microarray data. Bioinformatics 20:2726–2737

Mariadason JM, Nicholas C, L'Italien KE, Zhuang M, Smartt HJ et al (2005) Gene expression profiling of intestinal epithelial cell maturation along the crypt-villus axis. Gastroenterology 128:1081–1088

Marken JS, Schieven GL, Hellstrom I, Hellstrom KE, Aruffo A (1992) Cloning and expression of the tumor-associated antigen L6. Proc Natl Acad Sci USA 89:3503–3507

Martin F, Penet MF, Malergue F, Lepidi H, Dessein A et al (2004) Vanin-1(-/-) mice show decreased NSAID- and Schistosoma-

induced intestinal inflammation associated with higher glutathione stores. J Clin Invest 113:591–597
McCole DF, Barrett KE (2007) Varied role of the gut epithelium in mucosal homeostasis. Curr Opin Gastroenterol 23:647–654
Moehle C, Ackermann N, Langmann T, Aslanidis C, Kel A et al (2006) Aberrant intestinal expression and allelic variants of mucin genes associated with inflammatory bowel disease. J Mol Med 84:1055–1066
Mutch DM, Simmering R, Donnicola D, Fotopoulos G, Holzwarth JA et al (2004) Impact of commensal microbiota on murine gastrointestinal tract gene ontologies. Physiol Genomics 19:22–31
Nakajima D, Nakayama M, Kikuno R, Hirosawa M, Nagase T et al (2001) Identification of three novel non-classical cadherin genes through comprehensive analysis of large cDNAs. Brain Res Mol Brain Res 94:85–95
Niimi T, Nagashima K, Ward JM, Minoo P, Zimonjic DB et al (2001) Claudin-18, a novel downstream target gene for the T/EBP/NKX2.1 homeodomain transcription factor, encodes lung- and stomach-specific isoforms through alternative splicing. Mol Cell Biol 21:7380–7390
Okochi H, Mine T, Nashiro K, Suzuki J, Fujita T et al (1999) Expression of tetraspans transmembrane family in the epithelium of the gastrointestinal tract. J Clin Gastroenterol 29:63–67
Rahner C, Mitic LL, Anderson JM (2001) Heterogeneity in expression and subcellular localization of claudins 2, 3, 4, and 5 in the rat liver, pancreas, and gut. Gastroenterology 120:411–422
Reichling T, Goss KH, Carson DJ, Holdcraft RW, Ley-Ebert C et al (2005) Transcriptional profiles of intestinal tumors in Apc(Min) mice are unique from those of embryonic intestine and identify novel gene targets dysregulated in human colorectal tumors. Cancer Res 65:166–176
Scholzel S, Zimmermann W, Schwarzkopf G, Grunert F, Rogaczewski B et al (2000) Carcinoembryonic antigen family members CEACAM6 and CEACAM7 are differentially expressed in normal tissues and oppositely deregulated in hyperplastic colorectal polyps and early adenomas. Am J Pathol 156:595–605
Shi L, Reid LH, Jones WD, Shippy R, Warrington JA et al (2006) The MicroArray Quality Control (MAQC) project shows inter- and intraplatform reproducibility of gene expression measurements. Nat Biotechnol 24:1151–1161
Stegmann A, Hansen M, Wang Y, Larsen JB, Lund LR et al (2006) Metabolome, transcriptome, and bioinformatic cis-element analyses point to HNF-4 as a central regulator of gene expression during enterocyte differentiation. Physiol Genomics 27:141–155
Sugiyama Y, Farrow B, Murillo C, Li J, Watanabe H et al (2005) Analysis of differential gene expression patterns in colon cancer and cancer stroma using microdissected tissues. Gastroenterology 128:480–486
Tashima S, Shimada S, Yamaguchi K, Tsuruta J, Ogawa M (2000) Expression of brain-type glycogen phosphorylase is a potentially novel early biomarker in the carcinogenesis of human colorectal carcinomas. Am J Gastroenterol 95:255–263
Todd SC, Doctor VS, Levy S (1998) Sequences and expression of six new members of the tetraspanin/TM4SF family. Biochim Biophys Acta 1399:101–104
Tytgat KM, Buller HA, Opdam FJ, Kim YS, Einerhand AW et al (1994) Biosynthesis of human colonic mucin: Muc2 is the prominent secretory mucin. Gastroenterology 107:1352–1363
Verstege MI, te Velde AA, Hommes DW (2006) Apoptosis as a therapeutic paradigm in inflammatory bowel diseases. Acta Gastroenterol Belg 69:406–412
Wehkamp J, Salzman NH, Porter E, Nuding S, Weichenthal M et al (2005) Reduced Paneth cell alpha-defensins in ileal Crohn's disease. Proc Natl Acad Sci USA 102:18129–18134
Weiber H, Lindstrom C, Lilja H, Bjartell A, Fernlund P (1997) Immunohistochemical and in situ hybridization studies of beta-microseminoprotein in the human gastric mucosa. Histochem J 29:839–845
Weiber H, Borch K, Sundler F, Fernlund P (1999) Beta-microseminoprotein in gastric carcinoids: a marker of tumour progression. Digestion 60:440–448
Wright MD, Ni J, Rudy GB (2000) The L6 membrane proteins—a new four-transmembrane superfamily. Protein Sci 9:1594–1600
Yoshikai Y (1999) The interaction of intestinal epithelial cells and intraepithelial lymphocytes in host defense. Immunol Res 20:219–235