An Expression Profile of Active Genes in Human Colonic Mucosa

Kousaku Okubo,1* Junji Yoshii,1,2 Hideoki Yokouchi,3 Masao Kameyama,3 and Kenichi Matusbara1

Institute for Molecular and Cellular Biology, Osaka University 1-3, Yamada-oka, Suita, Osaka 565, Japan, 1
Hitachi Software Engineering, Co., Ltd. 6-81, Onoe-chou, Naka-ku, Yokohama 231, Japan, 2 and
Department of Surgery, Center for Adult Diseases Osaka 1-3-3, Nakamachi, Higashinari, Osaka 537, Japan3

(Received 2 December 1993)

Abstract

An expression profile of genes active in the human colonic mucosa was obtained by collecting 959 partial sequences from a 3'-directed cDNA library. Seven genes were found to produce mRNA each of which comprised more than 1% of total mRNA. Four of these genes are novel, and are likely to be uniquely expressed in the colonic mucosa, and the other three have been identified as genes for fatty acid binding protein, immunoglobulin lambda chain, and carcinoma-associated antigen GA733-2. In the remaining 952 clones, 310 were composed of 118 species occurred recurrently but less than 1%, and 533 clones appeared only once. Because the 3'-directed cDNA library faithfully represents the mRNA population in the source tissue, these numbers represent the relative activities of the gene expression.

Altogether 156 gene species were identified in GenBank, and a significant portion of these genes encode proteins found in Golgi apparatus and lysosomes, chromosome-encoded mitochondrial proteins, cell surface proteins, and components in the protein synthesis machinery. The types and proportions of genes identified is consistent with the known major activities of the colonic mucosa such as mucous protein production, energy-dependent water absorption, and rapid cell proliferation and turnover.

Key words: cDNA sequencing; 3'-directed cDNA library; gene signatures; novel genes; tissue specificity

1. Introduction

Gene expression profiles are a compilation of the names of all active genes in any given cell type or tissue and the relative amounts of the gene transcripts in the total mRNA. We have already begun collecting expression profiles of active genes with as many human cells or tissues as are available with the aim of constructing a body map in which we map genes to the site(s) of the body where they are active. This approach is intended to complement the structural analyses of the human genome.

The expression profile of the active genes is determined by sequencing randomly selected clones in a 3'-directed cDNA library which is free from the potential bias associated with cloning efficiencies, normalization or size fractionation. The resulting short sequences just upstream of the poly(A) tail are termed gene signatures (GSs), and are queried to the GenBank and compared within the library to identify the gene and to measure the recurrence, or abundance, of the cDNA.

In previous works, we prepared expression profiles of active genes with a cultured liver cell line (HepG2),1 a promyelocyte cell line (HL60), and the granulocytoid and monocytoid cells [manuscript in preparation] induced from the HL60. Here, we report the expression profile of the human colonic mucosa. The major functions of the colonic mucosa are to absorb large amounts of water and secrete large amounts of mucous substances. Due to the harsh conditions in the digestive gut, the cells survive only a limited period of time and undergo rapid turnover. The colonic mucosa consists mostly of goblet cells and absorptive cells. In addition, lymphoid cells and nodules with a large number of B cells are located in the loose connective tissue beneath the epithelium (lamina propria).

2. Materials and Methods

2.1. Colonic Mucosa

Mucosal tissue was peeled off from the marginal area of resected sigmoid colon of a 71-year-old male patient with primary colon cancer. This area was histologically proven to be negative for cancer invasion.

2.2. Construction of the 3'-Directed cDNA Library

Poly(A) RNA isolated from colonic mucosa (0.5 g) using the QuickPrep mRNA purification kit (Pharmacia
Uppsala, Sweden) was used for construction of the 3’-directed cDNA library in which the 3’ end fragment of MboI digested cDNAs were cloned in a PUC19-based vector primer.2

3. Sequencing and Data Analyses

A small portion of the ligation mixture was transformed into Escherichia coli DH5, and fresh transformant colonies were randomly selected, lysed and the cDNA insert moieties were amplified by polymerase chain reaction (PCR). The products were subjected to cycle sequencing using dye-labeled primers and analysed on Applied Biosystems Sequencers (Mountain View, CA, USA).3 From the sequence data, we eliminated clones with inserts shorter than 20 bp, clones having more than 5% ambiguous nucleotides, clones having mitochondrial DNA sequences, clones having Alu sequences immediately upstream of the poly(A) stretch. Sequences thus selected were then compared with each other to obtain the frequency of appearance in the library and then sent to GenBank (Re78) for gene identification using the FastA program.7 The 455 GSs newly registered in this work have been deposited in the DDBJ with locus names corresponding to gs numbers, such as HUMGS02546 for gsO2546, and with accession numbers D25548 through D26002.

4. Results and Discussion

mRNA isolated from the human colonic mucosa was used as a template for cDNA synthesis by vector-priming. The cDNA moiety was selectively cleaved by MboI, and a 3’-directed cDNA library was constructed and transformed into Escherichia coli. Although the average size of the inserts was about 250~270 bp, and a significant fraction of the clones lacked amino acid coding information, this type of library is unique in that its composition faithfully represents the mRNA population, and the short sequence data that represents the region immediately upstream of the poly(A) region can serve as a sequence tag for gene identification and to determine the abundance of gene transcripts in the mRNA population. We call these short sequences “gene signatures (GSs)”4 These GSs are different from the so-called expression sequence tags (ESTs) reported by Adams et al.5 that were obtained with a randomly primed cDNA library. With the randomly primed, non-directed cDNA library one can not tell whether two different ESTs represent two different mRNA species or whether they are derived from different parts of the same mRNA molecule. Normalized or so-called singlebook cDNA library for eliminating redundancy, and the size-fractionated cDNA library for enriching full size components are not suitable for quantitative analyses of the mRNA population.

We randomly chose transformant clones, then selectively amplified the cDNA moieties and sequenced them. Among those clones analyzed, we obtained 1161 GSs, of which 202 were identified as representing 14 species of mitochondrial transcripts. In the remaining 959 GSs, 116 GSs were attributed to 7 species that appeared 10 times or more and 310 GSs were attributed to 118 species that appeared recurrently but less than 10 times and 533 which appeared only once. Altogether, 658 species of GSs for chromosomally coded genes were represented in the collection (Table 1). The expression profile of active genes in the human colonic mucosa is shown in Table 2. For clarity’s sake only those genes whose cDNA clones appeared 3 times or more are listed in the table. Of the most redundant 7 species, genes for the fatty acid binding protein, immunoglobulin lambda chain, and carcino-associated antigen GA733-2 were identified in GenBank, and the other 4 were novel. It is interesting to note that the gene for fatty acid binding protein codes for liver type, not the intestinal type protein that is active in the small intestine.6 Although the implication of this difference awaits further study, its role in the absorption of fatty acids in the bowel is doubtful since the major function of the colon is absorption of water, not lipids and their derivatives. The presence of a significant number of B cells in the colonic mucosa is reflected by the immunoglobulin gene products, although at present we do not know why only immunoglobulin lambda chain gene products have been detected.

The expression profile of the genes in other cell types which we have analysed are listed in Table 2, including the liver-derived HepG2,1 the promyelocyte-derived HL60, and the granulocytoid and monocytoid cells each derived from HL60 by induction [manuscript in preparation]. By comparing the expression profiles, the above-mentioned four GSs representing novel genes (gs02546, gs04094, gs02123 and gs02706) are likely to appear only in the mRNA of the colonic mucosa. Although the number of cells and tissues we have analyzed is not large,
Table 2. Expression profile of active genes in colon mucosa.

| GS      | size | total | HepG2 | HL60 | granulo | mono | colon | sim% | sim size | definition                                                                 |
|---------|------|-------|-------|------|---------|------|-------|------|----------|-----------------------------------------------------------------------------|
| gso0196 | 282  | 41    | 1     |      |         |      |       | 40   | M10617   | 95    liver fatty acid binding protein (L-FABP)                              |
| gso2546 | 102  | 18    |       |      |         |      |       | 18   |          |                                              |
| gso4094 | 255  | 13    |       |      |         |      |       |      |          |                                              |
| gso2842 | 100  | 12    |       |      |         |      |       | 12   | M87790   | 98    immunoglobulin lambda chain                                            |
| gso2123 | 188  | 12    |       |      |         |      |       |      |          |                                              |
| gso2706 | 273  | 11    |       |      |         |      |       |      |          |                                              |
| gso4116 | 212  | 10    |       |      |         |      |       |      | M93036   | 97    212 carcinoma-associated antigen GA733-2                              |
| gso1190 | 281  | 8     | 1     |      |         |      |       | 7    |          |                                              |
| gso0650 | 273  | 28    | 14    | 2    | 6       |      |       | 6    | L06432   | 97    265 18S ribosomal protein (HKE3)                                       |
| gso0273 | 196  | 38    | 9     | 6    | 4       | 13   |       | 6    | L13806   | 95    192 translationally controlled tumor protein                          |
| gso0335 | 176  | 30    | 1     | 16   | 4       | 3    |       | 6    | M36072   | 100   165 ribosomal protein L7a (surf 3) large subunit                       |
| gso1809 | 258  | 9     |       |      |         |      |       | 3    | M58485   | 99    258 lysosomal membrane glycoprotein CD63                              |
| gso0381 | 133  | 15    | 1     | 2    | 6       |      |       | 6    |          |                                              |
| gso0155 | 349  | 17    | 1     | 3    | 8       |      |       | 8    |          |                                              |
| gso0565 | 421  | 16    | 4     | 2    | 5       |      |       | 5    | M11948   | 94    321 promyelocytic leukemia cell clones pH58                            |
| gso2271 | 301  | 5     |       |      |         |      |       | 5    | M17987   | 95    419 beta-2-microglobulin                                              |
| gso1766 | 335  | 11    |       |      |         |      |       | 11   | V00594   | 93    318 metallothionein                                                  |
| gso0211 | 304  | 49    | 8     | 24   | 2       | 10   |       | 10   | X67247   | 96    170 ribosomal protein S8                                              |
| gso0285 | 190  | 34    | 6     | 14   | 1       | 8    |       | 8    |          |                                              |
| gso0917 | 68   | 12    |       |      |         |      |       | 12   |          |                                              |
| gso1657 | 42   | 7     |       |      |         |      |       | 7    |          |                                              |
| gso0162 | 329  | 14    | 1     | 4    | 3       | 2    |       | 2    | M17887   | 99    312 acidic ribosomal phosphoprotein P2                                |
| gso0019 | 551  | 57    | 17    | 8    | 2       | 26   |       | 26   | X16869   | 95    479 elongation factor 1-alpha (clone CEF4)                            |
| gso0818 | 124  | 10    |       |      |         |      |       | 10   |          |                                              |
| gso1670 | 33   | 5     |       |      |         |      |       | 5    |          |                                              |
| gso1919 | 137  | 8     |       |      |         |      |       | 8    |          |                                              |
| gso2221 | 69   | 3     |       |      |         |      |       | 3    | L20636   | 91    69 28S ribosomal RNA pseudogene                                       |
| gso4070 | 351  | 3     |       |      |         |      |       | 3    | M15402   | 94    317 carcinoembryonic antigen                                          |
| gso0804 | 176  | 10    | 2     | 1    | 2       | 2    |       | 2    | M26880   | 99    175 ubiquitin                                                        |
| gso4322 | 417  | 3     |       |      |         |      |       | 3    | M86400   | 100   415 phospholipase A2                                                 |
| gso1404 | 160  | 5     |       |      | 2       |      |       | 2    | M39351   | 96    159 set gene                                                         |
| gso1244 | 266  | 7     |       |      | 3       | 1    |       | 1    | X13585   | 94    262 mitochondrial ubiquinone-binding protein                          |
| gso0244 | 163  | 20    |       |      | 2       | 14   | 1     | 14   | X63432   | 96    163 beta-actin                                                        |
| gso0550 | 56   | 15    | 6     | 4    | 1       | 1    |       | 1    |          |                                              |
| gso0689 | 237  | 7     | 1     | 1    | 2       |      |       | 2    |          |                                              |
| gso0932 | 52   | 5     | 1     | 1    | 1       |      |       | 1    |          |                                              |
| gso1630 | 52   | 4     |       |      | 1       |      |       | 4    |          |                                              |
| gso1794 | 289  | 4     |       |      | 1       |      |       | 4    |          |                                              |
| gso2073 | 332  | 3     |       |      | 1       |      |       | 3    |          |                                              |
| gso2177 | 109  | 3     |       |      | 1       |      |       | 3    |          |                                              |
| gso2673 | 326  | 3     |       |      | 1       |      |       | 3    |          |                                              |
| gso2801 | 180  | 3     |       |      | 1       |      |       | 3    |          |                                              |
| gso4167 | 125  | 3     |       |      | 1       |      |       | 3    |          |                                              |
| gso0586 | 55   | 3     |       |      | 1       |      |       | 3    |          |                                              |

Abundant GSs appearing three times or more in the 959 chromosomally-coded GSs of colonic mucosa are listed in the descending order of abundance, along with their size and abundance so far obtained with libraries from HepG2 cells(1), HL60 cell(2), and granulocytoid and monocytoid cells (manuscript in preparation), each induced from HL60 by DMSO and phorbester, respectively. Gene names identified in GenBank (Re79) are listed in the far-right column. Colored rows represent those GSs that appeared 3 times or more in the colonic mucosa, but not in other libraries.
Table 3. Expression profile of active genes in colon mucosa.

| Gene Set (GS) | Size | GeneBank | Genes | % Similarity | Gene Name |
|---------------|------|----------|-------|--------------|-----------|
| **Energy production** |      |          |       |              |           |
| GS001107      | 231  | J04501   | 98.2  | 223          | muscle glycogen synthase |
| GS004040      | 374  | M83088   | 96.2  | 365          | phosphoglucomutase 1 (PGM1) |
| GS004026      | 302  | X05236   | 95    | 301          | fibroblast aldolase A |
| GS003122      | 240  | J03544   | 96.6  | 236          | brain glycogen phosphorylase |
| GS000504      | 54   | M14328   | 98    | 51           | alpha enolase |
| GS00422       | 112  | M10036   | 91.4  | 105          | triosephosphate isomerase |
| GS000145      | 332  | M57949   | 90.6  | 234          | triose-phosphate isomerase (TPI) |
| **Cytoplasmic miscellany** |      |          |       |              |           |
| GS004322      | 417  | M86400   | 99.8  | 415          | phospholipase A2 |
| GS000272      | 197  | J05262   | 100   | 197          | farnesyl pyrophosphate synthetase |
| GS004098      | 248  | M65131   | 94.7  | 246          | methylmalonyl-CoA mutase (MCM) |
| GS004191      | 87   | J03058   | 100   | 81           | argininosuccinate lyase |
| GS006666      | 216  | J05016   | 91.6  | 215          | deoxyxycytidine kinase (dCK kinase) gene |
| GS003689      | 409  | M88006   | 93.4  | 411          | S-adenosylmethionine decarboxylase (AMD1) |
| GS001995      | 66   | X13710   | 96.8  | 63           | glutathione peroxidase |
| GS000586      | 396  | M18377   | 95.5  | 396          | inositol polyposphate 1-phosphatase |
| GS003115      | 432  | J03037   | 91.5  | 435          | carbonic anhydrase II |
| GS000196      | 282  | M10617   | 94.9  | 274          | liver fatty acid binding protein (L-FABP) |
| GS001107      | 351  | M75883   | 97.7  | 347          | sterol carrier protein 2 |
| GS000135      | 373  | M11147   | 99.1  | 336          | ferritin L chain |
| GS001766      | 335  | V00594   | 93.4  | 318          | metallothionein |
| GS001919      | 142  | J02763   | 99.2  | 125          | calcyclin |
| GS000851      | 108  | D00761   | 100   | 87           | calmodulin |
| GS001140      | 165  | X56976   | 93.3  | 164          | ubiquitin activating enzyme E1 |
| GS000304      | 176  | M26880   | 99.4  | 175          | ubiquitin |
| GS000716      | 199  | X56999   | 96.5  | 199          | ubiquitin-52 amino acid fusion protein |
| **Golgi and lysosome** |      |          |       |              |           |
| GS003227      | 388  | M13519   | 96.1  | 387          | N-acetyl-beta-glucosaminidase (HEXB) |
| GS004105      | 233  | X53578   | 97.4  | 233          | "alpha(1.3/1.4)fucosyltransferase " |
| GS006799      | 29   | M29877   | 100   | 28           | alpha-L-fucosidase |
| GS001765      | 339  | M84739   | 94.7  | 337          | Soegren autoantigen/calreticulin of ER |
| GS004061      | 325  | M88458   | 94.2  | 325          | "ELP-1, regulator of Golgi to ER traffic" |
| GS001533      | 332  | J01482   | 98.1  | 105          | lamp-lencoding lysosomal membrane glycoprotein-1 |
| GS001809      | 258  | M58485   | 98.8  | 258          | lysosomal membrane glycoprotein CD63 |
| GS      | size | HepG2 | HL60 | Granulocytoids | Monocytes | Colon mucosa | Acc. in GenBank | % similarity | size of sim. | gene name                                      |
|---------|------|-------|------|----------------|-----------|--------------|----------------|--------------|-------------|------------------------------------------------|
| GS002821 407 | 2    | X68277 | 94.1 | 404            | CL 100 protein tyrosine phosphatase |
| GS006709 371 | 1    | L15388 | 95.7 | 369            | G protein-coupled receptor kinase (GRK5) |
| GS001720 464 | 1    | M14043 | 94.8 | 461            | lipocortin II |
| GS004131 187 | 2    | M75099 | 97.3 | 186            | rapamycin- and FK506-binding protein |
| GS004134 173 | 1    | M75099 | 98.3 | 173            | rapamycin- and FK506-binding protein |
| GS000270 212 | 1    | X04526 | 95.2 | 209            | beta-subunit signal transducing proteins Gs/Gi |
| GS004076 415 | 1    | X53143 | 98.2 | 273            | YL8 ras-like protein |
| GS002959 81  | 1    | X02751 | 100  | 76             | N-ras and flanking regions |
| GS006774 278 | 1    | L05367 | 98.6 | 77             | NF1 neurofibromatosis 1 (NF1) |
| GS003867 157 | 1    | M81457 | 99.3 | 145            | calpain 1 light chain |
| GS000019 581 | 17   | X16869 | 94.6 | 479            | elongation factor 1-alpha (clone CEF4) |
| GS006650 273 | 14   | L06432 | 97.4 | 265            | 18S ribosomal protein (HKE3) sequence. |
| GS000077 459 | 1    | L20868 | 96.3 | 456            | ribosomal protein L4 (RPL4) |
| GS000262 220 | 2    | M17885 | 98.6 | 210            | acidic ribosomal phosphoprotein P0 |
| GS000163 329 | 1    | M17886 | 95.7 | 305            | acidic ribosomal phosphoprotein P1 |
| GS000162 329 | 1    | M17887 | 98.7 | 312            | acidic ribosomal phosphoprotein P2 |
| GS000512 46  | 2    | M22146 | 100  | 43             | ribosomal protein S4 (scar protein) |
| GS000335 176 | 16   | M36072 | 100  | 165            | ribosomal protein L7a (surf 3) large subunit |
| GS004090 267 | 1    | M88459 | 99.6 | 264            | ribosomal protein (RPS4Y) |
| GS000545 516 | 3    | M94314 | 93.4 | 439            | ribosomal protein L30 (logue of yeast rpl30) |
| GS000336 167 | 2    | S45214 | 97   | 167            | yeast ribosomal protein YL41 homolog |
| GS000123 356 | 2    | X55954 | 98.8 | 336            | HL23 ribosomal protein homologue |
| GS000519 39  | 3    | X63527 | 97.2 | 36             | ribosomal protein L19 |
| GS000290 184 | 5    | X66699 | 98.4 | 184            | ribosomal protein L37a |
| GS000807 133 | 2    | X69392 | 99.2 | 133            | ribosomal protein L26 |
| GS004144 159 | 2    | X69654 | 98.7 | 156            | ribosomal protein S26 |
| GS000583 382 | 2    | X73460 | 95.8 | 378            | ribosomal protein L3 |
| GS000211 304 | 8    | X67247 | 95.9 | 170            | ribosomal protein S8 |
| GS002271 301 | 5    | M34482 | 95.6 | 298            | cytokeratin 8 (CK8) |
| GS000243 227 | 1    | X12881 | 100  | 225            | cytokeratin 18 |
| GS000114 387 | 1    | X04098 | 92.2 | 383            | gamma-actin |
| GS000244 163 | 2    | X63432 | 95.7 | 163            | beta-actin |
| GS003121 410 | 2    | X02344 | 92.9 | 410            | beta-tubulin |
| GS001808 255 | 6    | S54005 | 98   | 253            | thymosin beta-10 |
| GS000258 220 | 1    | X04588 | 95.4 | 217            | cytoskeletal tropomyosin TM30(sm) |
| GS001154 308 | 2    | J03191 | 96.7 | 301            | profilin |
| GS000595 322 | 2    | L05491 | 98.4 | 320            | T-plastin (actin binding protein) |
| GS size HepG2 | Mitochondria | GS000755 132 | GS004165 154 | GS004259 144 | GS004089 270 | GS000801 135 | GS006480 43 | GS001695 526 | GS000683 247 | GS001244 266 | GS000799 141 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| size GenBank of sim. | Mitochondria | 132 | 154 | 144 | 270 | 135 | 43 | 526 | 247 | 266 | 141 |
| size name | 132 | 154 | 144 | 270 | 135 | 43 | 526 | 247 | 266 | 141 | |
| Gene name | fERG | tERG | ADP/ATP translocase | 112,554 | 106,541 | 114,552 | 98,564 | 102,551 | 96,543 | 100,552 | 98,564 |
| Nucleus | | | | | | | | | | | |
| GS000219 264 | GS000797 150 | GS001811 204 | GS000172 310 | GS002321 143 | GS000156 369 | GS004136 169 | GS000136 249 | GS000797 150 | GS001811 204 | GS000172 310 | GS002321 143 |
| size HepG2 | Mitochondria | GS000029 384 | GS000156 369 | GS000048 370 | GS000201 195 | GS000351 187 | GS000640 216 | GS000351 187 | GS000640 216 | GS000351 187 | GS000640 216 |
| size name | Proteins | 384 | 369 | 370 | 195 | 187 | 216 | 187 | 216 | 187 | 216 |
| Gene name | D13888 | J09894 | M11534 | J04718 | M28136 | J07507 | M12301 | J07507 | M12301 | J07507 | M12301 |
| Secretory | | | | | | | | | | | |
| GS003919 116 | GS000565 421 | GS002148 141 | GS002542 100 | GS000212 267 | GS000173 306 | GS002716 379 | GS003663 49 | GS000173 306 | GS002716 379 | GS003663 49 | |
| size HepG2 | Mitochondria | GS000029 384 | GS000156 369 | GS000048 370 | GS000201 195 | GS000351 187 | GS000640 216 | GS000351 187 | GS000640 216 | GS000351 187 | GS000640 216 |
| size name | Proteins | 384 | 369 | 370 | 195 | 187 | 216 | 187 | 216 | 187 | 216 |
| Gene name | L12350 | M36501 | M36501 | M26658 | X04412 | Y00264 | X04412 | Y00264 | X04412 | Y00264 | X04412 |

Gene Expression Profile in Human Colonic Mucosa

Vol. 1.
| GS          | size | HepG2 | HL60 | Granulocytoids | Monocytoids | Colon mucosa | GenBank acc. | %similarity | size of sim. | gene name                        |
|-------------|------|-------|------|---------------|-------------|--------------|--------------|-------------|-------------|----------------------------------|
| GS002931    | 106  |       | 1    | 1             | 1           | D10653       | 100          | 100         | cell surface glycoprotein        |
| GS001991    | 69   |       | 1    | 1             | 1           | L06132       | 98.5         | 66          | voltage-dependent anion channel 1 (VDAC) |
| GS004088    | 272  |       | 1    | 1             | 3           | M15042       | 94.3         | 317         | receptor tyrosine kinase         |
| GS004070    | 351  |       | 1    | 1             | 3           | M15042       | 94.3         | 317         | carcinoembryonic antigen          |
| GS004103    | 238  |       | 1    | 1             | 1           | X03945       | 96.6         | 237         | HLA-B37 antigen                   |
| GS004116    | 212  |       | 1    | 1             | 10          | M03036       | 97.2         | 212         | carcinoma-associated antigen GA733-2 (GA733-2) |
| GS003447    | 364  |       | 1    | 1             | 1           | S54769       | 96.2         | 234         | cellular adhesion regulatory molecule |
| GS002656    | 357  |       | 1    | 1             | 2           | X00497       | 92.3         | 351         | HLA-DR antigens associated invariant chain (p33) |
| GS004057    | 332  |       | 1    | 1             | 1           | X02160       | 100          | 154         | insulin receptor precursor        |
| GS004110    | 239  |       | 1    | 1             | 1           | X03363       | 99.2         | 236         | c-erb-B-2                         |
| GS004003    | 39   |       | 1    | 1             | 1           | X06614       | 97.2         | 36          | receptor of retinoic acid         |
| GS006714    | 231  |       | 1    | 1             | 1           | X15187       | 97.8         | 231         | tral homolog of murine tumor rejection antigen gp96 |
| GS006747    | 183  |       | 1    | 1             | 1           | Y00815       | 98.8         | 172         | LCA-homolog, LAR protein (leukocyte antigen related) |
| GS000501    | 215  |       | 1    | 1             | 1           | L00352       | 93.6         | 218         | low density lipoprotein receptor  |
| GS001933    | 278  |       | 1    | 1             | 1           | X60111       | 97.5         | 276         | motility related protein (MRP-1) / CD9 |
| GS004092    | 257  |       | 1    | 1             | 2           | M35252       | 98.8         | 250         | CO-029 tumor antigen              |
| GS003896    | 433  |       | 1    | 1             | 1           | D10522       | 96.1         | 407         | “80K-L protein”                   |
| GS001775    | 390  |       | 1    | 1             | 1           | D11094       | 96.3         | 374         | MSS1 Tat mediated transactivation modulator |

**Surface membrane**

| GS          | size | HepG2 | HL60 | Granulocytoids | Monocytoids | Colon mucosa | GenBank acc. | %similarity | size of sim. | gene name                        |
|-------------|------|-------|------|---------------|-------------|--------------|--------------|-------------|-------------|----------------------------------|
| GS004091    | 263  |       | 1    | 1             | 2           | L08044       | 92.2         | 234         | intestinal trefoil factor expressed in goblet cell |
| GS001388    | 166  |       | 1    | 1             | 2           | L10342       | 99.4         | 163         | novel gene 5' to ribonucleotide reductase M1 |
| GS002073    | 196  |       | 9    | 6             | 4           | L13806       | 94.8         | 192         | translationally controlled tumor protein |
| GS000155    | 349  |       | 1    | 3             | 8           | M11948       | 93.8         | 321         | promyelocytic leukemia cell clones pH58 |
| GS001960    | 114  |       | 1    | 1             | 1           | M35252       | 98.8         | 250         | CO-029 tumor antigen              |
| GS006764    | 337  |       | 1    | 1             | 1           | M18217       | 91.4         | 336         | DNA of undetermined origin found 5' to NCA |
| GS000865    | 100  |       | 9    | 2             | 1           | M55409       | 100          | 89         | pancreatic tumor-related protein  |
| GS004742    | 148  |       | 1    | 1             | 1           | M73547       | 95.9         | 148         | DP1 (deleted in poliposis coli) gene |
| GS000704    | 219  |       | 4    | 1             | 2           | M77024       | 92.5         | 133         | cross-reactive protein to ARF sera |
| GS000078    | 378  |       | 1    | 1             | 1           | M80783       | 97.6         | 375         | B12 (TNF inducible endothelial protein) |
| GS006379    | 296  |       | 1    | 1             | 1           | M83751       | 94.9         | 296         | arginine-rich gene                 |
| GS001404    | 160  |       | 2    | 1             | 3           | M03651       | 96.2         | 159         | set gene                          |
| GS004049    | 346  |       | 1    | 1             | 1           | S45936       | 95.7         | 350         | HTS1=HeLa tumor suppressor gene   |
| GS001967    | 95   |       | 1    | 1             | 1           | X14986       | 100          | 57         | t-complex TCP1 gene homologue     |
| GS002712    | 398  |       | 1    | 1             | 1           | X53331       | 94.1         | 392         | matrix Gl protein                 |
| GS000503    | 52   |       | 1    | 2             | 1           | X57346       | 98.1         | 52         | HS1 protein                       |

**Uncharacterized**

"80K-L protein"
Table 3. Continued

| GS       | size HepG2 | HL60 Granulocytes | Monocytoids | Colon Acc. in mucosa | GenBank | %similarity | size of sim. | gene name                                      |
|----------|------------|-------------------|-------------|----------------------|---------|-------------|-------------|-----------------------------------------------|
| GS000102 | 391        | 2                 | 4           | 3                    | X64707  | 93.3        | 387         | BBC1                                         |
| GS002059 | 358        | 1                 |             | X67698               | 97.4    | 346         | tissue specific epididymus mRNA              |
| GS003698 | 352        | 2                 |             | M19645               | 96.9    | 351         | 78 kdalton glucose-regulated protein (GRP78) |
| GS00723  | 201        | 1                 |             | L10376               | 97      | 202         | (clone CTG-B33) human mRNA with triplet repeat (Li et al.) |
| GS003864 | 144        | 1                 |             | L13803               | 99.3    | 143         | liver expressed protein fragment (Bonald et al) |
| GS00464  | 86         | 1                 | 1           | D13630               | 98.6    | 74          | randomly sampled cDNA clone (deposited by Nomura et al) |
| GS00422  | 485        | 1                 |             | D13641               | 93.5    | 431         | randomly sampled cDNA clone (deposited by Nomura et al) |
| GS004100 | 245        | 1                 |             | D14658               | 100     | 243         | human ORF (deposited by Nomura et al)         |
| GS003062 | 394        | 1                 |             | D14662               | 97.1    | 375         | human ORF (deposited by Nomura et al)         |
| GS000555 | 481        | 1                 |             | D14662               | 97.5    | 481         | human ORF (deposited by Nomura et al)         |
| GS000363 | 144        | 1                 |             | D14696               | 98.6    | 141         | human ORF with identical sequence to HepG2 mRNA |

GSs identified in GenBank (Re78) are listed according to their function or subcellular locality. From left to right, each of the column represents GS number, size of the GS, represented frequency of the GS among about 1000 isolates from 3' directed library of HepG2(1), HL60, granulocytoids, monocytoids (manuscript in preparation), accession number of corresponding sequence in GenBank, similarity percent, size of the similar region (bp) between GS and GenBank entry found by FastA program (7) and GenBank definition.
this observation suggests that these four novel genes may code for products that are unique to the colonic mucosa and that are produced in relatively large amounts. In contrast, protein synthesis genes that are ubiquitous are active in many different cell types (Table 2). Among the other GSs whose abundance was less than 10 in Table 2, ten appeared only in the colonic mucosa mRNA. They are also likely to represent genes that act only in the colonic mucosa.

Genes for protein synthesis are unusually active in the colonic mucosa, and at least 7 such genes are listed in the top 44 in Table 2. We have observed similar properties with HepG2 and HL60, both of which are actively replicating cells. In contrast, granulocyte cells that do not divide, show significantly lowered expression of such genes [manuscript in preparation]. Epithelial cells in the colonic mucosa are known to be replaced rapidly by the newly proliferated cells and the elevated expression of genes related to protein synthesis may reflect this unique property.

The 156 identified genes, representing 307 clones (Table 1), are grouped according to their functions and cellular locations, and are listed in Table 3. In this list we can see several clones that could be expected based on the unique structure and function of the colonic mucosa. First, production of large amounts of mucus proteins is reflected by the number of clones for proteins in Golgi apparatus. Second, there are significant numbers of clones for glycolysis enzymes and mitochondrial proteins, a pattern that has not been observed with the HepG2 liver cell line. Active energy production may be coupled with the active pumping of water into the body by this tissue. The presence of a large number of mitochondria has been well known with the absorptive cells in the colon mucosa. Third, there are a number of clones for surface membrane proteins and signal transducers. In addition, there are clones for voltage-dependent anion channel, colon mucosa-associated DRA, intestinal trefoil factor of goblet cells, and the DP1 gene. Each one of these gene products plays a characteristic role in the colonic mucosa. Also listed are genes for cytokeratin 8 and 18, which are typically found in epithelial cells with well-developed desmosomes. On the other hand, the sodium pump gene has not yet been identified in the colon. Whether this gene(s) is expressed at a low level, or whether it is one of the novel genes is not yet known.

To our knowledge, this is the first systematic analysis of active genes in human colonic mucosa. Of all the genes whose clones were identified, 60% of those which appeared twice or more in this experiment are novel, reflecting the paucity of biochemical data with this tissue. As a natural consequence of these studies, detailed characterizations of such novel genes are needed, especially those whose products are likely to be specific to the colonic mucosa. As we have shown previously, random sequencing of about 1000 clones is sufficient to determine the expression profile of the genes in any given tissue, whether, ubiquitous or cell-specific, that are abundantly expressed in the tissue. On the other hand, to examine genes that are expressed at low levels, we must avoid recurring sequences of the same transcript.

The entire intestine is divided into several regions, such as the duodenum, jejunum, ileum, cecum, colon, and rectum, each of which play different roles in the body. Extending this work to the mucosa from these different regions will enrich our understanding of these tissues at the molecular level, and at the same time will enable us to identify many interesting novel genes whose functions are closely associated with the unique biological activities.

Acknowledgments: This work was supported in part by a Grant-in-Aid for Creative Basic Research (Human Genome Program) from the Ministry of Education, Science and Culture, Japan.

References

1. Okubo, K., Hori, N., Matoba, R., Niiyama, T., Fukushima, A., Kojima Y., and Matsubara, K. 1992, Large scale cDNA sequencing for analysis of quantitative and qualitative aspects of gene expression, Nature [Genetics], 2, 173-179.
2. Okubo, K., Hori, N., Matoba, R., Niiyama, T., and Matsubara, K. 1991, A novel system for large-scale sequencing of cDNA by PCR amplification, DNA Seq., 2, 137-44.
3. Hunkapiller, T., Kaiser, R. J., Koop, B. F., and Hood, L. 1991, Large-Scale and Automated DNA Sequence Determination, Science, 254, 59-67.
4. Matsubara, K. and Okubo, K. 1993, cDNA analysis in the human genome project, Gene, 135, 265-274.
5. Adams, M. D., Kelley, J. M., Gocayne, J. D., Dubnick, G. M., Polymeropoulos, M. H., Xiao, H., Merrill, C. R., Wu, A., Olde, B., Moreno, R. F. et al. 1991, Complimentary DNA Sequencing: Expressed sequence Tags and the human genome project, Science, 252, 1651-1656.
6. Gordon, J. I. and Lowe, J. B. 1985, Analyzing the structures, functions and evolution of two abundant gastrointestinal fatty acid binding proteins with recombinant DNA and computational techniques, Chem. Phys. Lipids, 38, 137-58.
7. Pearson, W. R. and Lipman, D. J. 1988, Improved tools for biological sequence comparison, Proc. Natl. Acad. Sci. U.S.A., 85, 2444-2448.
