Journal of Carcinogenesis

Special Issue - Colorectal Carcinogenesis and Prevention

Original Article

Systematic enrichment analysis of gene expression profiling studies identifies consensus pathways implicated in colorectal cancer development

Jesús Lascorz[1]*, Kari Hemminki[1,2], Asta Försti[1,2]

[1]Division of Molecular Genetic Epidemiology, German Cancer Research Center (DKFZ), Heidelberg, Germany. [2]Center for Primary Health Care Research, Clinical Research Center, Lund University, Malmö, Sweden.

E-mail: j.lascorz@dkfz.de
*Corresponding author

Published: 24 March, 2011
Journal of Carcinogenesis 2011, 10:7
This article is available from: http://www.carcinogenesis.com/content/10/1/7

Received: 14 December, 2010
Accepted: 22 January, 2011

Abstract

Background: A large number of gene expression profiling (GEP) studies on colorectal carcinogenesis have been performed but no reliable gene signature has been identified so far due to the lack of reproducibility in the reported genes. There is growing evidence that functionally related genes, rather than individual genes, contribute to the etiology of complex traits. We used, as a novel approach, pathway enrichment tools to define functionally related genes that are consistently up- or down-regulated in colorectal carcinogenesis. **Materials and Methods:** We started the analysis with 242 unique annotated genes that had been reported by any of three recent meta-analyses covering GEP studies on genes differentially expressed in carcinoma vs normal mucosa. Most of these genes (218, 91.9%) had been reported in at least three GEP studies. These 242 genes were submitted to bioinformatic analysis using a total of nine tools to detect enrichment of Gene Ontology (GO) categories or Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways. As a final consistency criterion the pathway categories had to be enriched by several tools to be taken into consideration. **Results:** Our pathway-based enrichment analysis identified the categories of ribosomal protein constituents, extracellular matrix receptor interaction, carbonic anhydrase isozymes, and a general category related to inflammation and cellular response as significantly and consistently overrepresented entities. **Conclusions:** We triaged the genes covered by the published GEP literature on colorectal carcinogenesis and subjected them to multiple enrichment tools in order to identify the consistently enriched gene categories. These turned out to have known functional relationships to cancer development and thus deserve further investigation.

Keywords: Carcinogenesis, colorectal cancer, enrichment analysis, gene expression profiling

Access this article online

Quick Response Code:

Website: www.carcinogenesis.com

DOI: 10.4103/1477-3163.78268

BACKGROUND

Colorectal cancer (CRC) is the third most common cancer, comprising 9.7% of all cancer cases, and is the fourth leading cause of cancer death worldwide, accounting for 8% of all cancer deaths.[1] Many gene expression profiling (GEP) studies on colorectal carcinogenesis have been performed

in the last decade using microarray technology. However, comparative analysis of the differentially expressed genes reported by independent studies shows a relatively limited degree of overlap, and no reliable biomarker profile discriminating cancerous from normal tissue has been identified. The majority of the published GEP studies on colorectal carcinogenesis has already been subjected to meta-analyses that have aimed at establishing consistent signature profiles for tumor development.[2–4] These meta-analyses have collected published lists of differentially expressed genes from the original GEP studies comparing CRC to normal tissue and then selected the genes reported in multiple studies. The genes reported only sporadically are thought to have resulted from inherent noise or biases in the different platforms and analysis methods employed.[5] The consistently reported genes are considered to be biologically relevant to CRC.

There is an increasing interest in searching for networks of genes, instead of single genes, contributing to the etiology of complex diseases, since changes in biological characteristics require coordinate variation in expression of gene sets.[6] Enrichment analysis tools, which estimate overrepresentation of particular gene categories or pathways in a gene list, are a useful approach in this direction.

Our goal was to define functional categories [Gene Ontology (GO) terms or Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways] that are consistently overrepresented among differentially expressed genes inferred from the published GEP studies on colorectal carcinogenesis. We collected the list of genes from three published meta-analyses and used them as an input list for an overrepresentation analysis with several independent enrichment tools, which are based on diverse statistical and bioinformatic algorithms.[7] The strategy of applying multiple tools is recommended for the most satisfactory results.[8] The stringent selection criteria for the genes to be analyzed and the requirement for concordance between enrichment analysis results helped us to identify consistently enriched gene categories of likely relevance in colorectal carcinogenesis.

MATERIALS AND METHODS

Gene expression profiling studies

We collected data from three meta-analyses, covering 34 GEP studies on the colorectal carcinogenesis process, published between the years 2001 and 2007.[2-4] Two of the meta-analyses reported a list of genes which had a consistent direction in gene expression change between carcinoma and normal mucosa in at least three single GEP studies,[2,3] while the threshold was two GEP studies in the oldest meta-analysis[4] [Table 1].

Gene list collection

For the meta-analysis by Sagynaliev *et al*.[4] we used Entrez Gene from NCBI (www.ncbi.nlm.nih.gov/gene/), and the Gene ID conversion tool from the DAVID bioinformatics resources[9] to convert the reported gene identifiers into the official HUGO gene symbol, which was used as the identifier for the reported genes. Next, the three gene lists from the three meta-analyses were combined, resulting in a list of 242 unique annotated genes [Table 2].

Enrichment analysis

We performed enrichment analyses using the databases GO (Biological Process and Molecular Function),[10] and KEGG pathways.[11] For all enrichment tools, the input gene set consisted of the same 242-gene list. The nine selected enrichment software tools differed in the statistical model applied for the enrichment analysis and in the method of correction for multiple testing [Table 3]. The tools were used with the default options: significance threshold of 0.05 for adjusted *P* value, at least two genes from the input list in the enriched category, and the whole genome as the reference background. For GATHER, the recommended ln(Bayes factor) >6 was used as the significance threshold.

Consistently enriched categories

We considered only the GO or KEGG categories reported to be significantly enriched by several enrichment tools as consistently overrepresented in the 242-gene list. This strategy, based on testing multiple tools, is recommended in order to obtain the most satisfactory results.[8] We selected as a threshold the number of tools reporting at least four common enriched categories, so that only top-ranked categories were finally considered. This threshold was five enrichment tools for GO Biological Process, six enrichment tools for GO Molecular Function, and three enrichment tools for KEGG pathways [Table 4].

Table 1: Three meta-analyses of gene expression profiling studies on CRC carcinogenesis process

First author	Ref.	Year	Number of GEP studies included	Selection discriminating genes	Number of reported discriminating mapped genes
Cardoso	[2]	2007	17	Reported by ≥3 independent studies	128
Chan	[3]	2008	23	Reported by ≥3 independent studies	163
Sagynaliev	[4]	2005	7#	Reported by ≥2 independent studies	68*

#Twelve studies were originally reported but five were not considered because they were performed either in samples from only two patients or in cell lines and not in patient samples. *Number of unique annotated mapped genes converted from the originally reported gene identifiers.

Table 2: The list of 242 unique, annotated genes reported in the three meta-analyses of GEP studies on CRC carcinogenesis used for enrichment analyses

Gene symbol	Name	Up/Down regulated in cancer vs. normal
ABP1	Amiloride binding protein 1 (amine oxidase (copper-containing))	down
ACAA2	Acetyl-Coenzyme A acyltransferase 2 (mitochondrial 3-oxoacyl-Coenzyme A thiolase)	down
ACADS	Acyl-Coenzyme A dehydrogenase, C-2 to C-3 short chain	down
ADH1A	Alcohol dehydrogenase 1A, α polypeptide	down
ADH1B	Alcohol dehydrogenase 1B (class I), beta polypeptide	down
ADH1C	Alcohol dehydrogenase 1C (class I), γ polypeptide	down
AHCYL2	Adenosylhomocysteinase-like 2	down
ANPEP	Alanyl (membrane) aminopeptidase	down
APBA3	Amyloid beta (A4) precursor protein-binding, family A, member 3	down
ATP5A1	ATP synthase, H+ transporting, mitochondrial F1 complex, alpha subunit 1, cardiac muscle	down
ATP5B	ATP synthase, H+ transporting, mitochondrial F1 complex, beta polypeptide	down
BCAS1	Breast carcinoma amplified sequence 1	down
C1ORF115	Chromosome 1 open reading frame 115	down
CA1	Carbonic anhydrase I	down
CA12	Carbonic anhydrase XII	down
CA2	Carbonic anhydrase II	down
CA4	Carbonic anhydrase IV	down
CA7	Carbonic anhydrase VII	down
CCL19	Chemokine (C-C motif) ligand 19	down
CCNYL1	Cyclin Y-like 1	down
CD177	CD177 molecule	down
CEACAM1	Carcinoembryonic antigen-related cell adhesion molecule 1 (biliary glycoprotein)	down
CEACAM7	Carcinoembryonic antigen-related cell adhesion molecule 7	down
CES2	Carboxylesterase 2 (intestine, liver)	down
CFD	Complement factor D (adipsin)	down
CGA	Glycoprotein hormones, alpha polypeptide	down
CHGA	Chromogranin A (parathyroid secretory protein 1)	down
CKB	Creatine kinase, brain	down
CLCA1	Chloride channel accessory 1	down
CLEC3B	C-type lectin domain family 3, member B	down
CLU	Clusterin	down
CNN1	Calponin 1, basic, smooth muscle	down
DGKH	Diacylglycerol kinase, eta	down
EDN3	Endothelin 3	down
ENTPD5	Ectonucleoside triphosphate diphosphohydrolase 5	down
FABP1	Fatty acid binding protein 1, liver	down
FCGBP	Fc fragment of IgG binding protein; similar to IgGFc-binding protein precursor (FcgammaBP)	down
FHL1	Four and a half LIM domains 1	down
FXYD3	FXYD domain containing ion transport regulator 3	down
GCG	Glucagon	down
GCNT3	Glucosaminyl (N-acetyl) transferase 3, mucin type	down
GPA33	Glycoprotein A33 (transmembrane)	down
GPX2	Glutathione peroxidase 2 (gastrointestinal)	down
GSN	Gelsolin (amyloidosis, Finnish type)	down
GUCA1B	Guanylate cyclase activator 1B (retina)	down
GUCA2A	Guanylate cyclase activator 2A (guanylin)	down
GUCA2B	Guanylate cyclase activator 2B (uroguanylin)	down
HMGCS2	3-hydroxy-3-methylglutaryl-Coenzyme A synthase 2 (mitochondrial)	down
HPGD	Hydroxyprostaglandin dehydrogenase 15-(NAD)	down
HSD11B2	Hydroxysteroid (11-beta) dehydrogenase 2	down
HSD17B2	Hydroxysteroid (17-beta) dehydrogenase 2	down
ITM2C	Integral membrane protein 2C	down
KLF4	Kruppel-like factor 4 (gut)	down
KRT17	Keratin 17	down
KRT20	Keratin 20	down

(Cond...)

Table 2: Contd....

Gene symbol	Name	Up/Down regulated in cancer vs. normal
KRT8	Keratin 8	down
LGALS3	Lectin, galactoside-binding, soluble, 3	down
LGALS4	Lectin, galactoside-binding, soluble, 4	down
LRMP	Lymphoid-restricted membrane protein	down
MALL	Mal, T-cell differentiation protein-like	down
MAPK3	Mitogen-activated protein kinase 3	down
MEP1A	Meprin A, alpha (PABA peptide hydrolase)	down
MGC27165	Hypothetical protein MGC27165	down
MGLL	Monoglyceride lipase	down
MS4A12	Membrane-spanning 4-domains, subfamily A, member 12	down
MT1A	Metallothionein 1A	down
MT1G	Metallothionein 1G	down
MT1H	Metallothionein 1H	down
MT2A	Metallothionein 2A	down
MUC12	Mucin 12, cell surface associated; similar to mucin 11	down
MUC2	Mucin 2, oligomeric mucus/gel-forming	down
MYH11	Myosin, heavy chain 11, smooth muscle	down
MYL9	Myosin, light chain 9, regulatory	down
MYLK	Myosin light chain kinase	down
NCAM2	Neural cell adhesion molecule 2	down
PGM1	Phosphoglucomutase 1	down
PLS1	Plastin 1 (I isoform)	down
PRDX6	Peroxiredoxin 6	down
PRKACB	Protein kinase, cAMP-dependent, catalytic, beta	down
PYY	Peptide YY	down
SECTM1	Secreted and transmembrane 1	down
SELENBP1	Selenium binding protein 1	down
SEPP1	Selenoprotein P, plasma, 1	down
SLC26A2	Solute carrier family 26 (sulfate transporter), member 2	down
SLC26A3	Solute carrier family 26, member 3	down
SLC4A4	Solute carrier family 4, sodium bicarbonate cotransporter, member 4	down
SMPDL3A	Sphingomyelin phosphodiesterase, acid-like 3A	down
SPIB	Spi-B transcription factor (Spi-1/PU.1 related)	down
SRI	Sorcin	down
SST	Somatostatin	down
STK39	Serine threonine kinase 39 (STE20/SPS1 homolog, yeast)	down
TMEM54	Transmembrane protein 54	down
TSPAN1	Tetraspanin 1	down
TSPAN7	Tetraspanin 7	down
TST	Thiosulfate sulfurtransferase (rhodanese)	down
UGT1A6	UDP glucuronosyltransferase 1 family, polypeptide A9	down
VIPR1	Vasoactive intestinal peptide receptor 1	down
ABHD2	Abhydrolase domain containing 2	up
AHCY	Adenosylhomocysteinase	up
APOA1	Apolipoprotein A-I	up
AZGP1	Alpha-2-glycoprotein 1, zinc-binding pseudogene 1	up
BGN	Biglycan	up
BMP4	Bone morphogenetic protein 4	up
BMP7	Bone morphogenetic protein 7	up
BST2	NPC-A-7; bone marrow stromal cell antigen 2	up
C2	Complement component 2	up
CBX3	Similar to chromobox homolog 3; chromobox homolog 3 (HP1 gamma homolog, Drosophila)	up
CCNB1	Cyclin B1	up
CCNB2	Cyclin B2	up
CCT3	Chaperonin containing TCP1, subunit 3 (gamma)	up
CCT6A	Chaperonin containing TCP1, subunit 6A (zeta 1)	up

(Cond...)

Table 2: Contd....

Gene symbol	Name	Up/Down regulated in cancer vs. normal
CCT7	Chaperonin containing TCP1, subunit 7 (eta)	up
CD44	CD44 molecule (Indian blood group)	up
CD46	CD46 molecule, complement regulatory protein	up
CD81	CD81 molecule	up
CDC25B	Cell division cycle 25 homolog B (S. pombe)	up
CDH3	Cadherin 3, type 1, P-cadherin (placental)	up
CDK10	Cyclin-dependent kinase 10	up
CDKN3	Cyclin-dependent kinase inhibitor 3	up
CFB	Complement factor B	up
CKS2	CDC28 protein kinase regulatory subunit 2	up
CLDN2	Claudin 2	up
COL11A1	Collagen, type XI, alpha 1	up
COL1A1	Collagen, type I, alpha 1	up
COL1A2	Collagen, type I, alpha 2	up
COL3A1	Collagen, type III, alpha 1	up
COL4A1	Collagen, type IV, alpha 1	up
CPNE1	Copine 1	up
CSE1L	CSE1 chromosome segregation 1-like (yeast)	up
CTSH	Cathepsin H	up
CXCL1	Chemokine (C-X-C motif) ligand 1 (melanoma growth stimulating activity, alpha)	up
CXCL2	Chemokine (C-X-C motif) ligand 2	up
CXCL3	Chemokine (C-X-C motif) ligand 3	up
CXCL9	Chemokine (C-X-C motif) ligand 9	up
DPEP1	Dipeptidase 1 (renal)	up
EEF1A1	Eukaryotic translation elongation factor 1, alpha 1	up
EIF2S2	Eukaryotic translation initiation factor 2, subunit 2 beta, 38kDa	up
EIF3A	Eukaryotic translation initiation factor 3, subunit A	up
EIF3B	Eukaryotic translation initiation factor 3, subunit B	up
EIF3E	Eukaryotic translation elongation factor 3, subunit E	up
ENC1	Ectodermal-neural cortex (with BTB-like domain)	up
ETV4	Ets variant 4	up
FCGR3A	Fc fragment of IgG, low affinity IIIa, receptor (CD16a)	up
FN1	Fibronectin 1	up
FPR2	Formyl peptide receptor 2	up
GAPDH	Glyceraldehyde-3-phosphate dehydrogenase	up
GARS	Glycyl-tRNA synthetase	up
GDF15	Growth differentiation factor 15	up
GGH	Gamma-glutamyl hydrolase (conjugase, folylpolygammaglutamyl hydrolase)	up
GNB2L1	Guanine nucleotide binding protein (G protein), beta polypeptide 2-like 1	up
GPX4	Glutathione peroxidase 4 (phospholipid hydroperoxidase)	up
GSTP1	Glutathione S-transferase pi 1	up
GTF3A	General transcription factor IIIA	up
H19	H19, imprinted maternally expressed transcript (non-protein coding)	up
HMGA1	Hypothetical LOC100130009; high mobility group AT-hook 1	up
HMGB1	High-mobility group box 1	up
HMGB2	High-mobility group box 2	up
HNRNPA1	Heterogeneous nuclear ribonucleoprotein A1-like 3	up
HNRNPH1	Heterogeneous nuclear ribonucleoprotein H1 (H)	up
HOMER1	Homer homolog 1 (Drosophila)	up
HSP90AA1	Heat shock protein 90kDa alpha (cytosolic), class A member 1	up
HSP90AB1	Heat shock protein 90kDa alpha (cytosolic), class B member 1	up
HSPD1	Heat shock 60kDa protein 1 (chaperonin)	up
HSPE1	Heat shock 10 kDa protein 1 (chaperonin 10)	up
IFITM1	Interferon induced transmembrane protein 1 (9-27)	up
IFITM2	Interferon induced transmembrane protein 2 (1-8D)	up
IMPDH1	IMP (inosine monophosphate) dehydrogenase 1	up

(Cond...)

Table 2: Contd....

Gene symbol	Name	Up/Down regulated in cancer vs. normal
IMPDH2	IMP (inosine monophosphate) dehydrogenase 2	up
INHBA	Inhibin, beta A	up
ITGA2	Integrin, alpha 2 (CD49B, alpha 2 subunit of VLA-2 receptor)	up
LCN2	lipocalin 2	up
LDHB	Lactate dehydrogenase B	up
MCM3	Minichromosome maintenance complex component 3	up
MIF	Macrophage migration inhibitory factor (glycosylation-inhibiting factor)	up
MMP1	Matrix metallopeptidase 1 (interstitial collagenase)	up
MMP11	Matrix metallopeptidase 11 (stromelysin 3)	up
MMP12	Matrix metallopeptidase 12 (macrophage elastase)	up
MMP3	Matrix metallopeptidase 3 (stromelysin 1, progelatinase)	up
MMP7	Matrix metallopeptidase 7 (matrilysin, uterine)	up
MYBL2	V-myb myeloblastosis viral oncogene homolog (avian)-like 2	up
MYC	V-myc myelocytomatosis viral oncogene homolog (avian)	up
NAP1L1	Nucleosome assembly protein 1-like 1	up
NEK2	NIMA (never in mitosis gene a)-related kinase 2	up
NME1	Non-metastatic cells 1, protein (NM23A)	up
NOS2	Nitric oxide synthase 2A (inducible)	up
NPM1	Nucleophosmin (nucleolar phosphoprotein B23, numatrin)	up
ODC1	Ornithine decarboxylase 1	up
PABPC1	Poly(A) binding protein, cytoplasmic 1	up
PCNA	Proliferating cell nuclear antigen	up
PLA2G16	Phospholipase A2, group XVI	up
POLR1D	Polymerase (RNA) I polypeptide D, 16kDa	up
PPIB	Peptidylprolyl isomerase B (cyclophilin B)	up
PRKDC	Similar to protein kinase, DNA-activated, catalytic polypeptide; protein kinase, DNA-activated, catalytic polypeptide	up
PYCR1	Pyrroline-5-carboxylate reductase 1	up
RAN	RAN, member RAS oncogene family	up
RBM12	RNA binding motif protein 12	up
RPL18A	Ribosomal protein L18	up
RPL23	Ribosomal protein L23	up
RPL29	Ribosomal protein L29	up
RPL3	Ribosomal protein L3; similar to 60S ribosomal protein L3 (L4)	up
RPL30	Ribosomal protein L30	up
RPL31	Ribosomal protein L31	up
RPL6	Ribosomal protein L6	up
RPL7	Ribosomal protein L7	up
RPL8	Ribosomal protein L8	up
RPLP2	Ribosomal protein, large, P2	up
rpmD	50S ribosomal protein L30	up
RPS18	Ribosomal protein S18	up
RPS19	Ribosomal protein S19	up
RPS2	Ribosomal protein S2	up
RPS23	Ribosomal protein S23	up
RPS5	Ribosomal protein S5	up
RPS7	Ribosomal protein S7	up
RPSA	Ribosomal protein SA	up
RRM2	Ribonucleotide reductase M2 polypeptide	up
S100A9	S100 calcium binding protein A9	up
S100P	S100 calcium binding protein P	up
SERPINE1	Serpin peptidase inhibitor, clade E (nexin, plasminogen activator inhibitor type 1), member 1	up
SLC12A2	Solute carrier family 12 (sodium/potassium/chloride transporters), member 2	up
SLC3A2	Solute carrier family 3 (activators of dibasic and neutral amino acid transport), member 2	up
SND1	Staphylococcal nuclease and tudor domain containing 1	up
SNRPB	Small nuclear ribonucleoprotein polypeptides B and B1	up

(Cond...)

Table 2: Contd....

Gene symbol	Name	Up/Down regulated in cancer vs. normal
SORD	Sorbitol dehydrogenase	up
SOX4	SRY (sex determining region Y)-box 4	up
SOX9	SRY (sex determining region Y)-box 9	up
SPARC	Secreted protein, acidic, cysteine-rich (osteonectin)	up
SPP1	Secreted phosphoprotein 1	up
STC1	Stanniocalcin 1	up
SULF1	Sulfatase 1	up
TACSTD2	Tumor-associated calcium signal transducer 2	up
TGFBI	Transforming growth factor, beta-induced, 68kDa	up
TGIF1	TGFB-induced factor homeobox 1	up
THBS2	Thrombospondin 2	up
TKT	Transketolase	up
TOMM40	Translocase of outer mitochondrial membrane 40 homolog (yeast)	up
TOP2A	Topoisomerase (DNA) II alpha 170kDa	up
TRAP1	TNF receptor-associated protein 1	up
TRIM28	Tripartite motif-containing 28	up
UBE2C	Ubiquitin-conjugating enzyme E2C	up
VEGFA	Vascular endothelial growth factor A	up
VSNL1	Visinin-like 1	up
WEE1	WEE1 homolog (S. pombe)	up

Table 3: Enrichment tools used and their characteristics

Tool name§	First reference	Databases	Key statistical method	Multiple testing correction method(s)
ConsensusPathDB	[20]	KEGG	Hypergeometric	FDR
DAVID	[9]	BP/MF/KEGG	EASE score (Fisher exact)	Benjamini*/FDR/Bonferroni
FatiGO	[21]	BP/MF	Fisher exact	3 methods (including B-H)
GATHER	[22]	BP/KEGG	Bayes factor	FDR
GeneCodis	[23]	BP/MF/KEGG	Hypergeometric	FDR
GOTM	[24]	BP/MF	Hypergeometric	B-H
g:Profiler	[25]	BP/MF/KEGG	Hypergeometric	g:SCS threshold
ToppFun	[26]	BP/MF/KEGG	Hypergeometric	Bonferroni/FDR*
WebGestalt	[27]	BP/MF/KEGG	Hypergeometric	B-H

BP: Gene ontology biological process; MF: Gene ontology molecular function; KEGG: Kyoto encyclopedia of genes and genomes; FDR: False discovery rate; B-H: Benjamini-Hochberg. *Indicates the multiple testing correction method used if more than one method possible.
URLs: ConsensusPathDB: http://cpdb.molgen.mpg.de/, DAVID: http://david.abcc.ncifcrf.gov/home.jsp, FatiGO: http://babelomics3.bioinfo.cipf.es/, GATHER: http://gather.genome.duke.edu/, GeneCodis: http://genecodis.dacya.ucm.es/, GOTM: http://bioinfo.vanderbilt.edu/gotm/, g: Profiler: http://biit.cs.ut.ee/gprofiler/index.cgi, ToppFun: http://toppgene.cchmc.org/ToppGene/enrichment.jsp, WebGestalt: http://bioinfo.vanderbilt.edu/wg_gsat/

Table 4: Number of overrepresented GO and KEGG categories in the 242-gene list for each of the enrichment tools used

Tool name	GO biological process	GO molecular function	KEGG pathways
ConsensusPathDB	n.a.	n.a.	2
DAVID	30	10	1
FatiGO	8	8	n.a.
GATHER	6	n.a.	0
GENECODIS	139	48	37
GOTM	6	3	n.a.
g:Profiler	48	16	4
ToppFun	22	14	2
WebGestalt	40	40	57
Significant categories ≥2 tools	47	30	36
Significant categories ≥3 tools	30	14	4*
Significant categories ≥4 tools	20	13	2
Significant categories ≥5 tools	10*	8	2
Significant categories ≥6 tools	3	5*	1

Only categories significantly ($P<.05$) enriched after correction for multiple testing is shown. GO: Gene ontology; KEGG: Kyoto encyclopedia of genes and genomes. n.a.: database not applicable. A threshold of at least four common enriched categories* was used to select the consistently enriched categories.

RESULTS

Data collection and gene selection

A total of 242 unique mapped genes [Table 2] were reported in at least one of the three meta-analyses (65 of them in two and 26 in all three meta-analyses), 145 (59.9%) of the genes were up-regulated and 97 (40.1%) down-regulated in cancer *vs* normal tissue. Twenty-four of the 242 genes (9.9%) had been reported by two single GEP studies and 218 genes (90.1%) by at least three single GEP studies.

Enrichment analyses

Nine enrichment tools were used to obtain significantly overrepresented categories (GO Biological Process, GO Molecular Function, and KEGG pathways) [Table 5].

Identification of consistently enriched categories

The number of reported enriched categories showed considerable variability with the different tools used [Table 4] even though the same significance threshold ($P < .05$ after correction for multiple testing) and analysis conditions (whole genome as the reference background and at least two genes from the input list in the enriched category) were applied. Differences were also observed in the number of genes in a particular category and the enrichment *P* values reported by each tool [Table 5]. To avoid false positives among the varying results, only the categories reported to be enriched by several tools (five enrichment tools for GO Biological Process, six for GO Molecular Function, and three for KEGG pathways) were considered to be consistently enriched. Using this selection criteria, ten general GO Biological Process categories (*cell proliferation*, *inflammatory response*, *multicellular organismal metabolic process*, *regulation of cell proliferation*, *response to chemical stimulus*, *response to external stimulus*, *response to nutrient*, *response to stress*, *response to wounding*, and *translational elongation*); five GO Molecular Function categories (*carbonate dehydratase activity*, *cytokine activity*, *extracellular matrix binding*, *receptor binding*, and *structural constituent of ribosome*); and four KEGG pathways (*extracellular matrix receptor interaction*, *focal adhesion*, *nitrogen metabolism*, and *ribosome*) were consistently overrepresented in the 242 gene list [Table 6]. The ratio of enrichment was higher for the more specific and well-defined KEGG pathways than for the broad GO categories [Figure 1]. A very high overlap of the individual genes among these categories was also observed [Table 7]. Based on this overlap, four biologically meaningful category groups were finally obtained:

a) Seventeen common genes included in the GO Biological Process *translational elongation*, the GO Molecular Function *structural constituent of ribosome*, and the KEGG pathway *ribosome*.
b) Genes in the two KEGG pathways *extracellular matrix receptor interaction* and *focal adhesion* that were also included in the

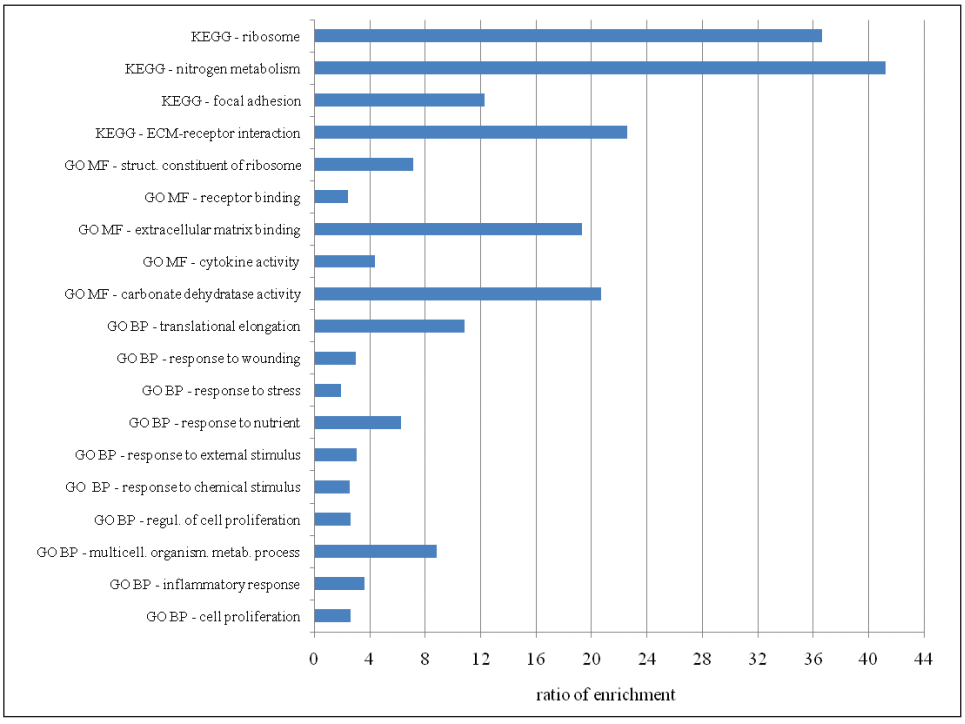

Figure 1: Bar chart of enrichment ratios for GO and KEGG categories in the 242-gene list. Ratio of enrichment = the number of observed genes divided by the number of expected genes from each GO or KEGG category in the 242-gene list (according to WebGestalt or, alternatively, DAVID or GOTM tools). GO BP: Gene Ontology Biological Process; GO MF: Gene Ontology Molecular Function; KEGG: Kyoto Encyclopedia of Genes and Genomes.

Table 5A: Results of all enrichment tools used with the 242 gene list: Gene ontology biological process categories

ID	Category	GOTM	Gather	WebGestalt	ToppFun	FatiGO	g:Profiler	DAVID	Genecodis
Total number of significant categories		6	6	40	22	8	48	30	139
GO:0048856	Anatomical structure development						3.37E-06 60	2.50E-02 54	
GO:0006820	Anion transport		4.23E-05 13						6.94E-03 3
GO:0009058	Biosynthetic process		8.44E-05 35			1.69E-03 40	1.22E-08 55		
GO:0001568	Blood vessel development						2.89E-05 13		2.73E-02 3
GO:0008283	Cell proliferation	3.57E-07 44	1.41E-05 40	1.07E-06 43	0.00E+00 50	1.69E-03 26	1.43E-05 17	3.71E-03 19	1.03E-04 13
GO:0044249	Cellular biosynthetic process		8.44E-05 34				8.36E-08 51		
GO:0030574	Collagen catabolic process			1.03E-02 4					9.34E-04 4
GO:0030199	Collagen fibril organization			1.45E-02 4					1.87E-03 4
GO:0032963	Collagen metabolic process			2.00E-04 7	6.01E-03 7		1.11E-06 6	3.61E-03 6	
GO:0006956	Complement activation			1.21E-02 5					2.71E-02 2
GO:0050974	Detection of mechanical stimulus involved in sensory perception						2.57E-05 4	4.49E-02 4	
GO:0007586	Digestion				2.73E-02 10	1.69E-03 8			1.11E-02 4
GO:0040007	Growth						3.80E-06 13	1.34E-02 11	2.82E-03 4
GO:0002376	Immune system process			5.00E-04 36	1.22E-03 41				
GO:0006954	Inflammatory response			8.76E-05 20	1.44E-03 21		8.63E-06 16	1.20E-03 17	9.76E-04 10
GO:0044419	Interspecies interaction between organisms			1.40E-02 13					1.13E-04 13
GO:0040011	Locomotion			7.10E-03 20			1.34E-05 18	8.77E-03 18	
GO:0008152	Metabolic process			7.10E-03 154			1.15E-11 134		4.84E-03 13
GO:0044259	Multicellular organismal macromolecule metabolic process			3.00E-04 7	9.73E-03 7		2.05E-06 6	5.33E-03 6	
GO:0044236	Multicellular organismal metabolic process			7.00E-04 7	4.32E-02 7	9.40E-03 4	5.88E-06 6	1.18E-02 6	
GO:0032501	Multicellular organismal process						8.78E-07 89	4.07E-02 80	
GO:0006730	One-carbon metabolic process		1.41E-05 7						9.29E-07 7
GO:0048513	Organ development						1.85E-07 50	9.20E-03 44	
GO:0048015	Phosphoinositide-mediated signaling			2.40E-03 8				1.33E-02 8	5.41E-05 6
GO:0008284	Positive regulation of cell proliferation			1.21E-02 16					2.63E-04 12
GO:0019538	Protein metabolic process			3.90E-03 70			9.98E-06 65		
GO:0065008	Regulation of biological quality			3.90E-03 42			2.13E-05 39	1.38E-02 37	
GO:0042127	Regulation of cell proliferation	7.30E-05 31		1.00E-04 30	3.80E-05 36		2.54E-08 32	5.24E-04 29	5.38E-03 5
GO:0002682	Regulation of immune system process			7.00E-04 18	7.63E-03 21			1.57E-02 16	

(Cond...)

Table 5A: Contd...

ID	Category	GOTM	Gather	WebGestalt	ToppFun	FatiGO	g:Profiler	DAVID	Genecodis
GO:0042221	Response to chemical stimulus	6.41E-06 53		2.71E-07 49	0.00E+00 59		1.84E-13 53	1.74E-07 48	
GO:0009719	Response to endogenous stimulus			7.00E-04 19	4.60E-03 21		6.34E-06 19	1.64E-03 19	
GO:0009605	Response to external stimulus	7.18E-07 45		1.68E-08 43	0.00E+00 48	7.69E-04 25	1.05E-11 41	5.15E-08 41	
GO:0009991	Response to extracellular stimulus			8.00E-05 15	3.36E-04 17		1.03E-07 16	5.85E-04 15	
GO:0009725	Response to hormone stimulus			7.00E-04 18	4.16E-03 20		5.73E-06 18	1.63E-03 18	
GO:0007584	Response to nutrient			8.00E-05 12	1.32E-04 14		9.68E-08 13	2.78E-04 13	1.80E-02 4
GO:0031667	Response to nutrient levels			8.76E-05 14	4.34E-04 16		1.59E-07 15	2.33E-04 15	
GO:0010033	Response to organic substance	8.07E-05 38		5.00E-04 27			3.52E-07 29	2.24E-03 26	
GO:0048545	Response to steroid hormone stimulus			3.90E-03 11	1.88E-02 13			1.97E-02 11	1.20E-05 5
GO:0050896	Response to stimulus			5.00E-04 83			4.42E-08 84	5.41E-04 78	
GO:0006950	Response to stress			2.00E-04 51		7.69E-04 34	4.45E-08 51	3.84E-04 48	6.97E-03 6
GO:0033273	Response to vitamin			5.00E-04 8	5.48E-03 9		1.14E-05 8	3.91E-03 8	
GO:0009611	Response to wounding			8.00E-05 26	4.16E-04 28		2.68E-07 24	9.99E-05 25	3.75E-02 3
GO:0001501	Skeletal system development						7.24E-06 16	1.12E-02 15	9.07E-03 6
GO:0043589	Skin morphogenesis						1.72E-05 3		8.58E-03 2
GO:0048731	System development						2.62E-07 60	1.34E-02 53	
GO:0006412	Translation			5.12E-06 24	1.20E-05 25		4.36E-07 21		8.58E-07 14
GO:0006414	Translational elongation	1.13E-10 18		4.99E-11 18	0.00E+00 18		5.68E-13 16		3.49E-14 16
Number of significant categories only in one tool		0	0	7	1	2	13	0	116

Only the categories selected by at least two enrichment tools are shown. In each case, the first row represents the overrepresentation *P* value adjusted for multiple testing, and the second row the number of genes in the category within the 242 gene list

Table 5B: Results of all enrichment tools used with the 242 gene list: Gene ontology molecular function categories

ID	Category	GOTM	WebGestalt	ToppFun	FatiGO	g:Profiler	DAVID	Genecodis
Total number of significant categories		3	40	14	8	16	10	48
GO:0004013	Adenosylhomocysteinase activity		5.10E-03 2					2.41E-03 2
GO:0005488	Binding		4.40E-03 207			1.17E-10 205		
GO:0005509	Calcium ion binding		8.70E-03 26			3.71E-05 28	9.69E-03 27	2.10E-04 17
GO:0004089	Carbonate dehydratase activity		2.00E-04 5	1.56E-03 5	8.29E-03 5	9.74E-07 5	7.38E-03 5	5.78E-06 5
GO:0043498	Cell surface binding		6.00E-03 4			3.27E-05 5		1.63E-03 3
GO:0008009	Chemokine activity		5.10E-03 5					7.72E-04 5
GO:0005518	Collagen binding		1.08E-02 4					1.79E-03 4
GO:0010853	Cyclase activator activity		1.80E-03 3	4.80E-02 3				
GO:0005125	Cytokine activity		2.00E-04 13	2.07E-02 14	6.71E-03 14	2.55E-06 13	1.25E-03 13	1.34E-04 9

(Cond...)

Table 5B: Contd...

ID	Category	GOTM	WebGestalt	ToppFun	FatiGO	g:Profiler	DAVID	Genecodis
GO:0050840	Extracellular matrix binding	1.97E-05 7	3.56E-06 7	5.00E-06 8		1.41E-07 7	5.10E-04 7	4.32E-08 6
GO:0005201	Extracellular matrix structural constituent		3.50E-03 7		9.75E-04 9		9.64E-03 8	4.00E-05 7
GO:0004602	Glutathione peroxidase activity		4.10E-03 3					1.12E-03 3
GO:0005539	Glycosaminoglycan binding		3.50E-03 9	3.15E-02 10				
GO:0030250	Guanylate cyclase activator activity		1.80E-03 3	4.80E-02 3				
GO:0008201	Heparin binding		6.00E-03 7					7.00E-04 7
GO:0005179	Hormone activity		2.90E-03 8				2.32E-02 8	1.44E-05 8
GO:0003938	IMP dehydrogenase activity		2.90E-03 2					1.08E-03 2
GO:0051287	NAD or NADH binding		5.00E-03 5			4.86E-06 7		2.70E-04 5
GO:0030235	Nitric-oxide synthase regulator activity		1.26E-02 2					6.58E-03 2
GO:0016614	Oxidoreductase activity, acting on CH-OH group of donors		2.00E-04 10	4.23E-03 10	6.12E-03 9	5.84E-06 10		
GO:0016616	Oxidoreductase activity, acting on the CH-OH group of donors, NAD or NADP as acceptor		2.00E-04 10	1.77E-03 10	7.06E-03	2.34E-06 10		
GO:0001871	Pattern binding		4.40E-03 9	2.87E-02 11				
GO:0048407	Platelet-derived growth factor binding		4.00E-04 4	9.22E-03 4		6.45E-06 4	1.93E-02 4	3.94E-05 4
GO:0005515	Protein binding		6.00E-04 152			2.03E-10 155	2.10E-02 130	5.27E-21 124
GO:0005102	Receptor binding		2.00E-04 31	2.53E-03 32	4.58E-03 26	3.11E-08 34	3.21E-04 31	4.96E-02 6
GO:0003723	RNA binding		4.40E-03 23					1.51E-06 21
GO:0003735	Structural constituent of ribosome	3.26E-05 18	3.01E-08 17	0.00E+00 17	1.81E-05 17	6.54E-07 15		4.92E-10 15
GO:0005198	Structural molecule activity	2.76E-06 38	7.07E-09 34	0.00E+00 34		3.56E-10 32	1.97E-02 21	
GO:0030911	TPR domain binding		5.10E-03 2					2.41E-03 2
GO:0051082	Unfolded protein binding		2.00E-04 10	3.11E-03 10	4.58E-03 10	4.46E-06 10		1.52E-06 10
Number of significant categories only in one tool		0	10	0	0	1	0	26

Only the categories selected by at least two enrichment tools are shown. In each case, the first row represents the overrepresentation *P* value adjusted for multiple testing, and the second row the number of genes in the category within the 242 gene list

Table 5C: Results of all enrichment tools used with the 242 gene list: KEGG pathway categories

ID	Category	Gather	WebGestalt	ConsensusPathDB	ToppFun	g:Profiler	DAVID	Genecodis
Total number of significant categories		0	57	2	2	4	1	37
KEGG330	Arginine and proline metabolism		6.38E-05 5					1.17E-03 5
KEGG5219	Bladder cancer		3.00E-04 4					3.79E-03 4
KEGG4110	Cell cycle		4.06E-06 8					1.70E-04 8
KEGG4062	Chemokine signaling pathway		3.00E-04 7					7.03E-03 7
KEGG270	Cysteine and methionine metabolism		2.20E-03 3					1.34E-02 3

(Cond...)

Table 5C: Contd...

ID	Category	Gather	WebGestalt	ConsensusPathDB	ToppFun	g:Profiler	DAVID	Genecodis
KEGG4060	Cytokine-cytokine receptor interaction		4.00E-04 8					1.10E-02 8
KEGG982	Drug metabolism - cytochrome P450		2.00E-04 5					1.41E-02 4
KEGG983	Drug metabolism - other enzymes		7.00E-04 4					6.03E-03 4
KEGG4512	ECM-receptor interaction		7.44E-10 10	1.44E-03 10	2.19E-02 10	2.89E-05 10	9.20E-03 10	1.11E-07 10
KEGG71	Fatty acid metabolism		2.14E-05 5					3.74E-03 4
KEGG4510	Focal adhesion		1.43E-09 13			6.10E-04 13		4.30E-07 13
KEGG480	Glutathione metabolism		3.13E-06 6					4.60E-03 4
KEGG10	Glycolysis / gluconeogenesis		8.03E-06 6					1.57E-03 5
KEGG4340	Hedgehog signaling pathway		7.20E-03 3					4.17E-02 3
KEGG980	Metabolism of xenobiotics by cytochrome P450		2.00E-04 5					1.40E-02 4
KEGG910	Nitrogen metabolism		2.03E-06 5			1.29E-04 5		3.34E-05 5
KEGG4621	NOD-like receptor signaling pathway		1.00E-04 5					1.67E-03 5
KEGG4114	Oocyte meiosis		7.20E-03 4					4.60E-02 4
KEGG4115	p53 signaling pathway		1.50E-03 4					1.45E-02 4
KEGG5200	Pathways in cancer		3.13E-06 12					2.96E-04 12
KEGG360	Phenylalanine metabolism		1.28E-02 2					4.89E-02 2
KEGG5020	PPAR signaling pathway		1.50E-03 4					1.40E-02 4
KEGG4914	Prion diseases		2.20E-03 3					1.34E-02 3
KEGG4914	Progesterone-mediated oocyte maturation		4.06E-06 7					1.62E-04 7
KEGG5215	Prostate cancer		3.40E-03 4					2.76E-02 4
KEGG230	Purine metabolism		7.00E-04 6					1.28E-02 6
KEGG240	Pyrimidine metabolism		4.30E-03 4					3.24E-02 4
KEGG4810	Regulation of actin cytoskeleton		2.90E-03 6					3.34E-02 6
KEGG830	Retinol metabolism		1.30E-03 4					4.96E-02 3
KEGG3010	Ribosome		3.84E-20 17	1.98E-08 17	0.00E+00 17	2.30E-09 15		5.77E-14 15
KEGG5222	Small cell lung cancer		4.00E-04 5					5.47E-03 5
KEGG4350	TGF-beta signaling pathway		4.92E-05 6					1.08E-03 6
KEGG350	Tyrosine metabolism		4.00E-04 4					2.76E-02 3
KEGG280	Valine, leucine and isoleucine degradation		4.00E-03 3					2.69E-02 3
KEGG4270	Vascular smooth muscle contraction		1.30E-03 5					1.49E-02 5
KEGG5110	Vibrio cholerae infection		7.20E-03 3					4.01E-02 3
Number of significant categories only in one tool		0	21	0	0	0	0	1

Only the categories selected by at least two enrichment tools are shown. In each case, the first row represents the overrepresentation *P* value adjusted for multiple testing, and the second row the number of genes in the category within the 242 gene list

Table 6: Consistently enriched GO and KEGG categories

ID	Category	Number of genes in category	Number of tools/number of genes*
GO Biological Process			
GO:0008283	Cell proliferation	1167	8 Tools 50 Genes
GO:0006954	Inflammatory response	380	5 Tools 21 Genes
GO:0044236	Multicellular organismal metabolic process	59	5 Tools 7 Genes
GO:0042127	Regulation of cell proliferation	854	6 Tools 36 genes
GO:0042221	Response to chemical stimulus	1521	5 Tools 59 Genes
GO:0009605	Response to external stimulus	669	6 Tools 48 Genes
GO:0007584	Response to nutrient	173	5 Tools 14 Genes
GO:0006950	Response to stress	1915	5 Tools 51 Genes
GO:0009611	Response to wounding	622	5 Tools 28 Genes
GO:0006414	Translational elongation	104	5 Tools 18 Genes
GO Molecular Function			
GO:0004089	Carbonate dehydratase activity	15	6 Tools 5 Genes
GO:0005125	Cytokine activity	203	6 Tools 14 Genes
GO:0050840	Extracellular matrix binding	29	6 Tools 8 Genes
GO:0005102	Receptor binding	944	6 Tools 34 Genes
GO:0003735	Structural constituent of ribosome	161	6 Tools 18 Genes
KEGG pathway			
KEGG4512	Extracellular matrix receptor interaction	58	6 Tools 10 Genes
KEGG4510	Focal adhesion	135	3 Tools 13 Genes
KEGG910	Nitrogen metabolism	75	3 Tools 5 Genes
KEGG3010	Ribosome	147	5 Tools 17 Genes

*In each case, the first row shows the number of enrichment tools reporting the category as significantly overrepresented and the second row shows the maximal number of genes from the category present in the input list of 242 genes.

broad categories of GO Molecular Function *receptor binding* and GO Biological Process *response to external stimulus*.

c) The five genes included in both the GO Molecular Function category *carbonate dehydratase activity* and the KEGG pathway *nitrogen metabolism*.

d) A large group of seven general GO Biological Process categories (*inflammatory response*, *response to chemical stimulus*, *response to external stimulus*, *response to nutrient*, *response to stress*, and *response to wounding*), together with two general GO Molecular Function categories (*cytokine activity* and *receptor binding*).

DISCUSSION

The large number of microarray studies on colorectal carcinogenesis has shown a low degree of overlap in the identified genes. We extracted the 242 unique genes reported in three meta-analyses of GEP studies on colorectal carcinogenesis.[2–4] Only the meta-analysis by Cardoso *et al*.[2] includes a descriptive exploration of the main GO categories present among the differentially expressed genes. In an attempt to overcome the known lack of reproducibility at individual gene level among the GEP studies, we used up to nine bioinformatic enrichment tools to statistically determine which GO categories or KEGG pathways were significantly overrepresented in the 242-gene list. A total of 34 independent GEP studies were included in the three meta-analyses. Most of them used whole-genome expression arrays, which include probes for expression analysis of thousands of genes. Thus, we used all genes in the genome as background for the enrichment analysis. Although this might be an overestimation, the heterogeneity in the

Table 7: Overlap of the genes from the consistently enriched GO and KEGG categories

ID	Category	Genes	GO:0008283	GO:0006954	GO:0044236	GO:0042127	GO:0042221	GO:0009605	GO:0007584	GO:0006950	GO:0009611	GO:0006414	GO:0004089	GO:0005125	GO:0050840	GO:0005102	GO:0003735	KEGG4512	KEGG4510	KEGG910	KEGG3010
			50	21	7	36	59	48	14	51	28	18	5	14	8	34	18	10	13	5	17
GO:0008283	Cell proliferation	50		6	3	36	20	16	6	12	8	1		6	4	13	1	1	2		1
GO:0006954	Inflammatory response	21			1	4	14	20	4	19	20			7	2	8		4	3		
GO:0044236	Multicell. organismal metabolic process	7				3	4	3	2	3	2				1	3		3	3		
GO:0042127	Regulation of cell proliferation	36					12	12	6	6	6			5	3	8		1	2		
GO:0042221	Response to chemical stimulus	59						35	14	35	20		2	10	2	18		5	6	2	
GO:0009605	Response to external stimulus	48							14	35	28	1		12	4	19	1	7	7		1
GO:0007584	Response to nutrient	14								7	5			3	2	7		4	3		
GO:0006950	Response to stress	51									23		1	10	3	14		5	5	1	
GO:0009611	Response to wounding	28												1	2	5		4	3		
GO:0006414	Translational elongation	18													1		17				17
GO:0004089	Carbonate dehydratase activity	5																		5	
GO:0005125	Cytokine activity	14													1	12	1	1	1		
GO:0050840	Extracellular matrix binding	8														5	1	3	4		1
GO:0005102	Receptor binding	34																6	7		
GO:0003735	Structural const. of ribosome	18																			17
KEGG4512	ECM-receptor interaction	10																	9		
KEGG4510	Focal adhesion	13																			
KEGG910	Nitrogen metabolism	5																			
KEGG3010	Ribosome	17																			

The number of genes from the 242-gene list belonging to each category is indicated as well as the number of overlapping genes between each pair of categories.

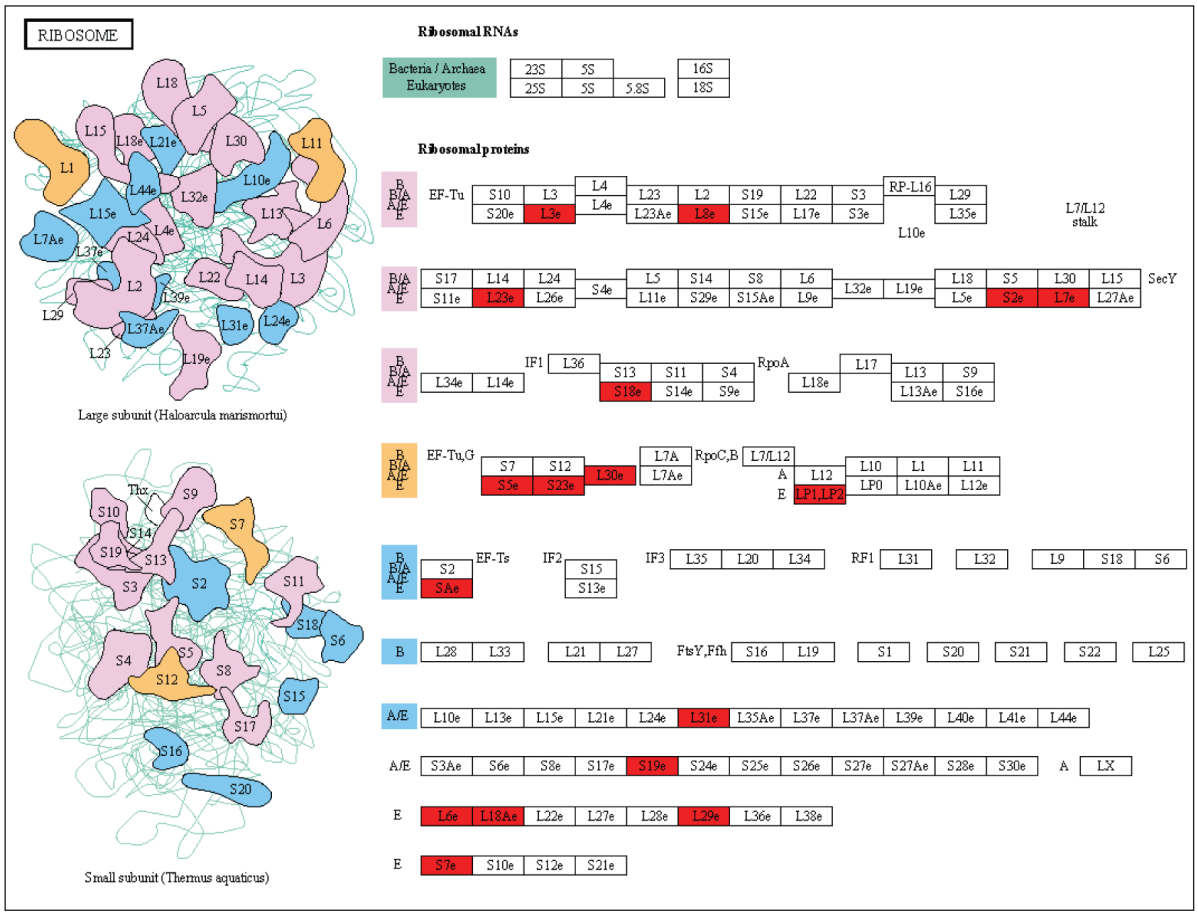

Figure 2: Representation of the KEGG ribosome category (map03010), with the 17 genes from the 242 gene list indicated in red

number of genes interrogated in every single one of the 34 GEP experiments does not allow application of a more appropriate restricted background. We believe that our rigorous strategy for the selection of enriched categories overcomes the forced probable overestimation of the reference background. After application of rigorous selection criteria, a total of 19 categories (15 GO terms and 4 KEGG pathways) were considered as consistently overrepresented. When considering the individual genes from each of these 19 categories, a very high degree of overlap among the categories was observed, reducing the number of categories with biological significance to four clearly different groups.

First, the same 17 ribosomal proteins (RPs) were present in the GO Biological Process *translational elongation*, the GO Molecular Function *structural constituent of ribosome*, and the KEGG pathway *ribosome* (RPL3, RPL6, RPL7, RPL8, RPL18A, RPL23, RPL29, RPL30, RPL31, RPLP2, RPSA, RPS2, RPS5, RPS7, RPS18, RPS19, and RPS23) [Figure 2]. All of them showed increased expression in tumor *vs* normal tissue. It is known that different expression patterns of RPs exist in CRC. Also, ribosomal biogenesis has clearly been linked to cancer[12] and several studies have pointed out two possible functions of RPs in colorectal carcinogenesis: perturbation of their function in protein biosynthesis and direct influence in tumorigenesis through extraribosomal functions (summarized in Lai *et al*.[13]). Second, the KEGG terms *extracellular matrix receptor interaction* and *focal adhesion* shared nine genes (COL1A1, COL1A2, COL3A1, COL4A1, COL11A1, FN1, ITGA2, SPP1, and THBS2) [Figure 3]. Specific interactions of the extracellular matrix molecules control cellular activities such as adhesion, differentiation, apoptosis, and proliferation.[14] Third, the GO category *carbonate dehydratase activity* and the KEGG pathway *nitrogen metabolism* included the same five carbonic anhydrase (CA) isozymes (CA1, CA2, CA4, CA7, and CA12) [Figure 4]. All five mRNAs are down-regulated in CRC compared to normal tissue, as also shown in another study for CA2 and CA12.[15] Recent data have confirmed the functional contribution of CAs, especially CA9 and CA12, to hypoxic tumor growth and progression.[16] Inhibition of CA9, which is overexpressed in many tumor types in response to the hypoxia inducible factor (HIF) pathway, is being tested as anticancer therapeutic strategy.[17] Finally, a very general

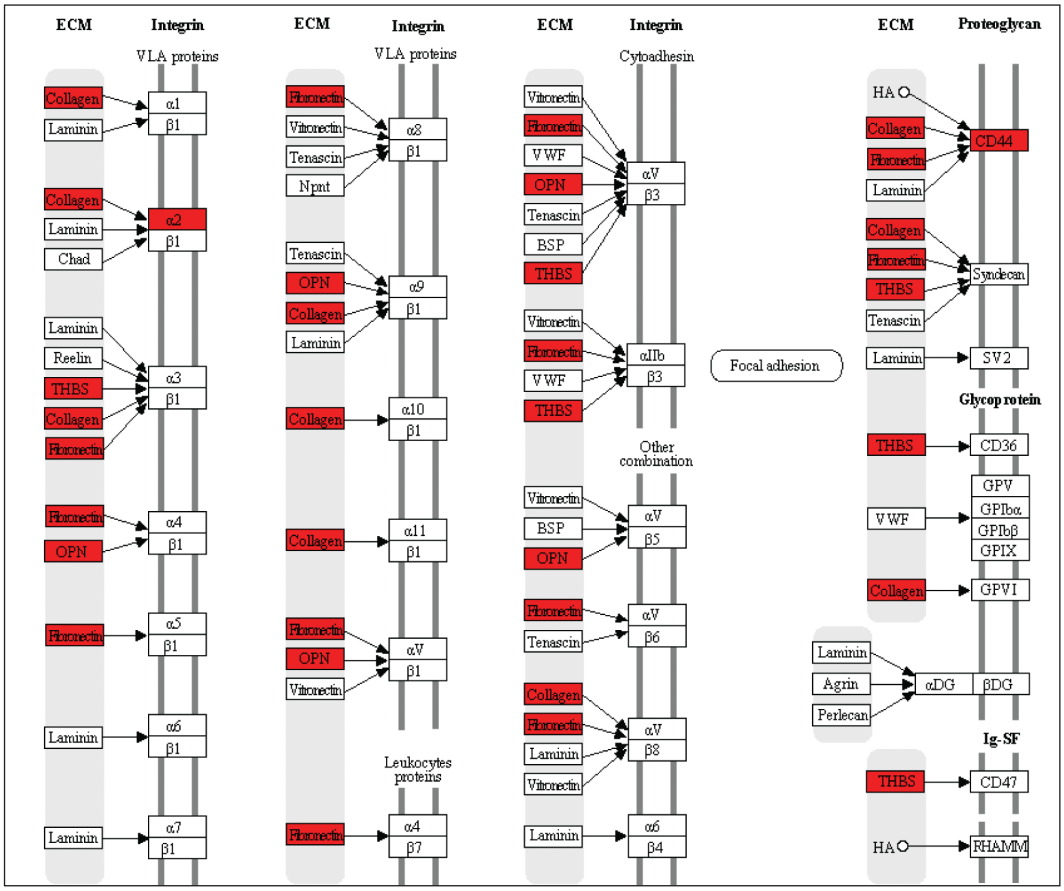

Figure 3: Representation of the KEGG extracellular matrix receptor interaction category (map04512), with location of the ten genes from the 242 gene list indicated in red.

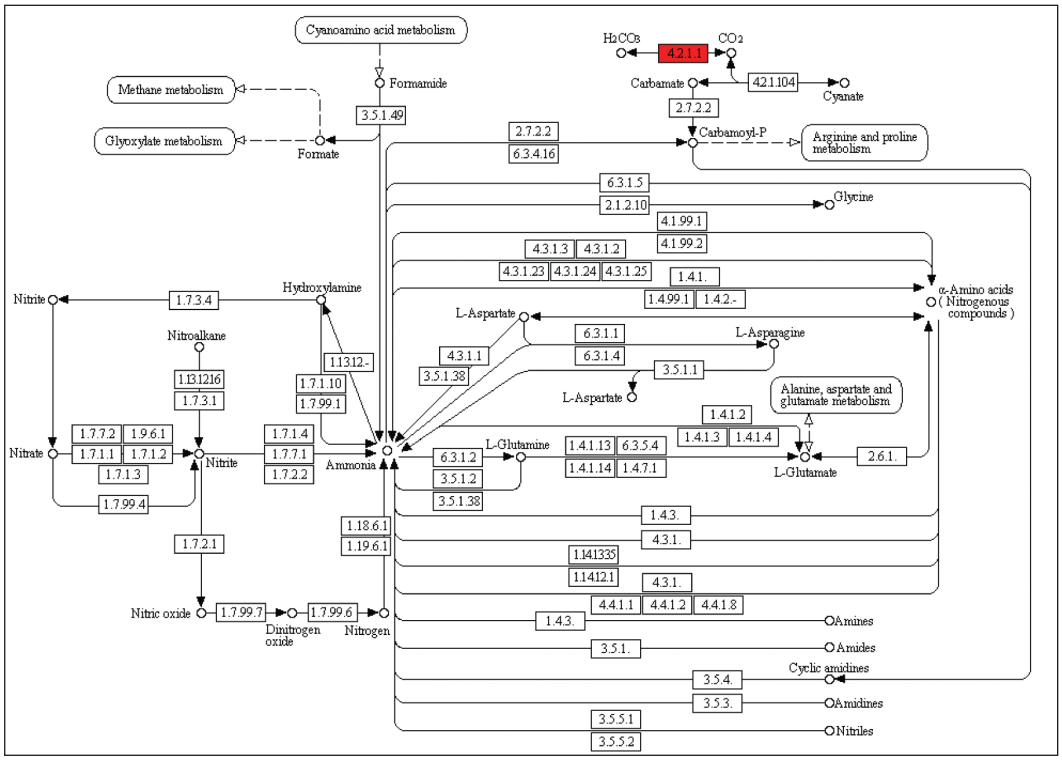

Figure 4: Representation of the KEGG nitrogen metabolism category (map00910), with location of the reaction catalyzed by the five carbonic anhydrase isozymes from the 242 gene list indicated in red

group of GO categories related to inflammation and cellular response included a large number of genes (between 14 and 59). Interestingly, this category included two genes that have been identified through genome-wide association studies as low-risk inherited genetic variants contributing to CRC risk.[18] These genes, the proto-oncogene MYC (8q24) and the bone morphogenetic protein gene BMP4 (14q22.2), were up-regulated in carcinoma tissue. Thus, judging by the functional class of the genes from the identified enriched categories, they look promising candidates for studies aimed at investigating their possible influence in CRC development.

In general, we observed a considerable variation in the number of enriched categories reported by each tool although there was uniformity in the analysis conditions used. However, despite this apparent variation, most of the enriched categories reported by the more stringent tools (those reporting a small number of enriched categories) were ranked among the top-categories by the more generous tools (those reporting a larger number of enriched categories). We considered this result of special interest because of previously reported lack of reproducibility between different enrichment tools.[7,8,19] This variability has been attributed to the statistical models applied by the enrichment analysis, to the method of correction for multiple testing, and to differences in the versions of the GO and KEGG data sources used. Thus, our strategy of using several bioinformatic tools to extract biologically related genes consistently involved in colorectal carcinogenesis proved to be successful.

CONCLUSIONS

We used the list of 242 unique mapped genes from three meta-analyses of GEP studies on colorectal carcinogenesis for a systematic enrichment analysis of GO categories and KEGG pathways, applying up to nine different enrichment tools. After applying stringent selection criteria to avoid false positive results, the ribosomal proteins group, the extracellular matrix receptor interaction category, the carbonic anhydrase isozymes, and a general category related to inflammation emerged as significantly and consistently overrepresented categories. These categories have known functional relationships to CRC development and their value as diagnostic markers and therapeutic targets deserve further investigation.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

JL and AF conceived and designed the study. JL conducted the analyses and wrote the initial manuscript. KH provided oversight and conceptual guidance to the project. KH and AF contributed to the final manuscript. All authors read and approved the final manuscript.

Funding

German National Genome Research Network (NGFN-Plus); the Deutsche Krebshilfe (German Cancer AID); and European Union (EU) [HEALTH-F4-2007-200767].

REFERENCES

1. Ferlay J, Shin HR, Bray F, Forman D, Mathers C, Parkin DM. Estimates of worldwide burden of cancer in 2008: GLOBOCAN 2008. Int J Cancer 2010 in press.
2. Cardoso J, Boer J, Morreau H, Fodde R. Expression and genomic profiling of colorectal cancer. Biochim Biophys Acta 2007;1775:103-37.
3. Chan SK, Griffith OL, Tai IT, Jones SJ. Meta-analysis of colorectal cancer gene expression profiling studies identifies consistently reported candidate biomarkers. Cancer Epidemiol Biomarkers Prev 2008;17:543-52.
4. Sagynaliev E, Steinert R, Nestler G, Lippert H, Knoch M, Reymond MA. Web-based data warehouse on gene expression in human colorectal cancer. Proteomics 2005;5:3066-78.
5. Siddiqui AS, Delaney AD, Schnerch A, Griffith OL, Jones SJ, Marra MA. Sequence biases in large scale gene expression profiling data. Nucleic Acids Res 2006;34:e83.
6. Hardy J, Singleton A. Genomewide association studies and human disease. N Engl J Med 2009;360:1759-68.
7. Huang da W, Sherman BT, Lempicki RA. Bioinformatics enrichment tools: Paths toward the comprehensive functional analysis of large gene lists. Nucleic Acids Res 2009;37:1-13.
8. Rhee SY, Wood V, Dolinski K, Draghici S. Use and misuse of the gene ontology annotations. Nat Rev Genet 2008;9:509-15.
9. Huang da W, Sherman BT, Lempicki RA. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat Protoc 2009;4:44-57.
10. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, *et al*. Gene ontology: Tool for the unification of biology: The Gene Ontology Consortium. Nat Genet 2000;25:25-9.
11. Kanehisa M, Goto S, Kawashima S, Okuno Y, Hattori M. The KEGG resource for deciphering the genome. Nucleic Acids Res 2004;32(Database issue):D277-80.
12. Lempiainen H, Shore D. Growth control and ribosome biogenesis. Curr Opin Cell Biol 2009;21:855-63.
13. Lai MD, Xu J. Ribosomal proteins and colorectal cancer. Curr Genomics 2007;8:43-9.
14. Desgrosellier JS, Cheresh DA. Integrins in cancer: Biological implications and therapeutic opportunities. Nat Rev Cancer 2010;10:9-22.
15. Niemela AM, Hynninen P, Mecklin JP, Kuopio T, Kokko A, Aaltonen L, *et al*. Carbonic anhydrase IX is highly expressed in hereditary nonpolyposis colorectal cancer. Cancer Epidemiol Biomarkers Prev 2007;16:1760-6.
16. Guler OO, De Simone G, Supuran CT. Drug design studies of the novel antitumor targets carbonic anhydrase IX and XII. Curr Med Chem 2010;17:1516-26.
17. Poulsen SA. Carbonic anhydrase inhibition as a cancer therapy: A review of patent literature, 2007 - 2009. Expert Opin Ther Pat 2010;20:795-806.
18. Tenesa A, Dunlop MG. New insights into the aetiology of colorectal cancer from genome-wide association studies. Nat Rev Genet 2009;10:353-8.
19. Khatri P, Draghici S. Ontological analysis of gene expression data: Current tools, limitations, and open problems. Bioinformatics 2005;21:3587-95.
20. Kamburov A, Wierling C, Lehrach H, Herwig R. ConsensusPathDB--a database for integrating human functional interaction networks. Nucleic Acids Res 2009;37(Database issue):D623-8.
21. Al-Shahrour F, Diaz-Uriarte R, Dopazo J. FatiGO: A web tool for finding significant associations of Gene Ontology terms with groups of genes. Bioinformatics 2004;20:578-80.
22. Chang JT, Nevins JR. GATHER: A systems approach to interpreting genomic signatures. Bioinformatics 2006;22:2926-33.

23. Carmona-Saez P, Chagoyen M, Tirado F, Carazo JM, Pascual-Montano A. GENECODIS:A web-based tool for finding significant concurrent annotations in gene lists. Genome Biol 2007;8:R3.
24. Zhang B, Schmoyer D, Kirov S, Snoddy J. GOTree Machine (GOTM): A web-based platform for interpreting sets of interesting genes using Gene Ontology hierarchies. BMC Bioinformatics 2004;5:16.
25. Reimand J, Kull M, Peterson H, Hansen J,Vilo J. g:Profiler--a web-based toolset for functional profiling of gene lists from large-scale experiments. Nucleic Acids Res 2007;35:W193-200.
26. Chen J, Bardes EE, Aronow BJ, Jegga AG. ToppGene Suite for gene list enrichment analysis and candidate gene prioritization. Nucleic Acids Res 2009;37:W305-11.
27. Zhang B, Kirov S, Snoddy J. WebGestalt: An integrated system for exploring gene sets in various biological contexts. Nucleic Acids Res 2005;33:W741-8.

AUTHOR'S PROFILE

Prof. Kari Hemminki, 1973 PhD in Medicine, University of Helsinki, Finland 1973 MD in Medicine, University of Helsinki, Finland 1975 Docent in Biochemistry, University of Helsinki, Finland 1976 - 1978 Postdoc in Molecular Biology, John Hopkins, Baltimore, USA

Jesus Lascorz, 2001 BsC in Biochemistry, University of Zaragoza, Spain 2002 - 2003 Molecular Genetics Laboratory, Central Institute of Mental Health, University of Heidelberg, Mannheim, Germany 2008 PhD in Human Biology, Institute of Human Genetics, University Erlangen-Nürnberg, Germany

Asta Foersti, 1984 M.Sc. in Biochemistry, University of Kuopio, Finland 1992 PhD in Biochemistry and Biotechnology, University of Kuopio, Finland