Open

Genes and Immunity (2011) 12, 653–662

www.nature.com/gene

ORIGINAL ARTICLE

CLC and IFNAR1 are differentially expressed and a global immunity score is distinct between early- and late-onset colorectal cancer

TH Ågesen[1,2], M Berg[1,2], T Clancy[3], E Thiis-Evensen[4], L Cekaite[1,2], GE Lind[1,2], JM Nesland[5,6], A Bakka[7], T Mala[8], HJ Hauss[9], T Fetveit[10], MH Vatn[4,11], E Hovig[3,12,13], A Nesbakken[2,6,8], RA Lothe[1,2,6] and RI Skotheim[1,2]

[1]*Department of Cancer Prevention, Institute for Cancer Research, The Norwegian Radium Hospital, Oslo University Hospital, Oslo, Norway;* [2]*Centre for Cancer Biomedicine, University of Oslo, Oslo, Norway;* [3]*Department of Tumor Biology, Institute for Cancer Research, The Norwegian Radium Hospital, Oslo University Hospital, Oslo, Norway;* [4]*Department for Organ Transplantation, Gastroenterology and Nephrology, Rikshospitalet, Oslo University Hospital, Oslo, Norway;* [5]*Division of Pathology, The Norwegian Radium Hospital, Oslo University Hospital, Oslo, Norway;* [6]*Faculty of Medicine, The University of Oslo, Oslo, Norway;* [7]*Department of Digestive Surgery, Akershus University Hospital, Lørenskog, Norway;* [8]*Department of Gastrointestinal Surgery, Aker, Oslo University Hospital, Oslo, Norway;* [9]*Department of Gastrointestinal Surgery, Sørlandet Hospital, Kristiansand, Norway;* [10]*Department of Surgery, Sørlandet Hospital, Arendal, Norway;* [11]*Epigen, Akershus University Hospital, Lørenskog, Norway;* [12]*Institute of Medical Informatics, The Norwegian Radium Hospital, Oslo University Hospital, Oslo, Norway and* [13]*Department of Informatics, The University of Oslo, Oslo, Norway*

Colorectal cancer (CRC) incidence increases with age, and early onset of the disease is an indication of genetic predisposition, estimated to cause up to 30% of all cases. To identify genes associated with early-onset CRC, we investigated gene expression levels within a series of young patients with CRCs who are not known to carry any hereditary syndromes (n = 24; mean 43 years at diagnosis), and compared this with a series of CRCs from patients diagnosed at an older age (n = 17; mean 79 years). Two individual genes were found to be differentially expressed between the two groups, with statistical significance; CLC was higher and IFNAR1 was less expressed in early-onset CRCs. Furthermore, genes located at chromosome band 19q13 were found to be enriched significantly among the genes with higher expression in the early-onset samples, including CLC. An elevated immune content within the early-onset group was observed from the differentially expressed genes. By application of outlier statistics, H3F3A was identified as a top candidate gene for a subset of the early-onset CRCs. In conclusion, CLC and IFNAR1 were identified to be overall differentially expressed between early- and late-onset CRC, and are important in the development of early-onset CRC.
Genes and Immunity (2011) **12,** 653–662; doi:10.1038/gene.2011.43; published online 30 June 2011

Keywords: *colorectal neoplasm; early onset of disease; gene expression microarray; genetic predisposition; hereditary cancer*

Introduction

The majority of colorectal cancers (CRCs) develop as sporadic disease, with incidence increasing with age.[1] The fraction of patients assumed to have an increased genetic risk accounts for 20–30% of all CRC cases, and early onset of disease is one indication of genetic predisposition.[2] However, less than 5% of the cases can be ascribed to known hereditary cancer syndromes, such as familial adenomatous polyposis and Lynch syndrome, with germline mutations in high-penetrance genes (*APC* and DNA mismatch-repair genes, respectively).[3]

Different strategies such as genetic linkage analysis,[4] DNA copy-number analysis,[5] gene expression analysis[6] and genome-wide association studies have been used to identify genetic factors that may predispose patients to cancer.[7] Interesting candidate genes have been highlighted, but the causal genetic variants underlying the increased risk remain to be identified. Recently, we identified novel CRC susceptibility loci containing potential oncogenes and tumor-suppressor genes by high-resolution microarray-based comparative genomic hybridization.[8] There have also been some attempts to reveal low-penetrance variants within genes that are already known to cause inherited CRC syndromes, focusing on genes in the DNA-repair pathway.[9,10]

Correspondence: Professor RA Lothe, Department of Cancer Prevention, Institute for Cancer Research, The Norwegian Radium Hospital, Oslo University Hospital, PO Box 4953 Nydalen, Oslo NO-0424, Norway.
E-mail: rlothe@rr-research.no
Received 17 March 2011; accepted 5 May 2011; published online 30 June 2011

Transcriptome analyses using microarray technology combined with different analytical approaches have been shown as a powerful strategy in the identification of genes that are associated with cancer development and progression, and to identify patterns related to clinical subgroups.[11–13] These data, in combination with advanced bioinformatic tools, have the potential to reveal changes in molecular pathways and to identify coherent expression of genes located in the same chromosomal region, in addition to single-gene expression differences. In the present study, we investigate the transcriptional differences between a potential genetic risk group of early-onset CRCs with no known hereditary cancer disease and a group of sporadic CRCs diagnosed at an older age.

Results

Characterization of the sample set and accompanying gene expression data

On using the most differentially expressed genes within the whole data set for a principal component analysis, the sample plot showed a mixed distribution of early- and late-onset tumors. The four normal colonic mucosa samples clustered together, distinct from the tumor samples (Figure 1a). Unsupervised hierarchical cluster analysis, using the same selection of differentially expressed genes, confirmed that the overall gene expression signatures were not distinct, neither between the different sample groups nor with regard to clinical parameters (Figure 1b).

The most differentially expressed genes between the cancer samples as a group and the normal colonic mucosa samples were further compared to published gene lists of similar comparisons.[14,15] A number of commonly reported genes with differential expression in cancer, regardless of age group, versus normal colonic mucosa, such as *TGFB1*, *CA2* and *MALL*, were in compliance with the results from our data (Supplementary Table 2), thus validating that the studied carcinomas were representative of CRC samples in general.

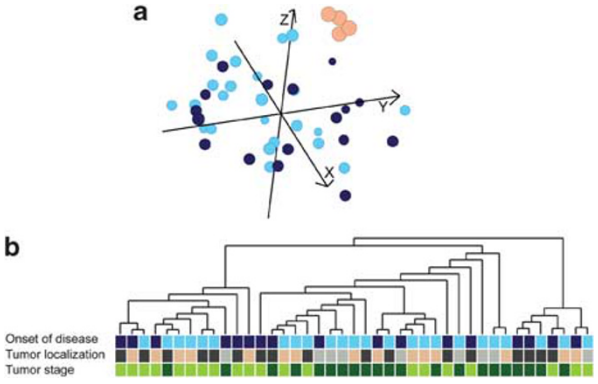

Figure 1 Early- and late-onset CRCs were not distinct with regard to their overall gene expression profiles. (**a**) The relatedness of gene expression profiles from 24 early-onset CRCs (light blue), 17 late-onset CRCs (dark blue) and four normal colonic mucosa (beige) samples as shown by their three first principal components. (**b**) A dendrogram of the CRC samples, resulting from hierarchical clustering analysis of gene expression data, illustrating sample distribution with regard to age (late onset, dark blue; early onset, light blue), clinical characteristics such as localization (left-sided tumors, dark gray; right-sided tumors (including transversum), gray; rectum, beige) and tumor stage (I and II, light green; III and IV, dark green).

Differentially expressed genes between early- and late-onset CRC

Two genes, *CLC* and *IFNAR1*, had statistically significant differential expression between the early- and late-onset CRCs (Significance Analysis of Microarrays (SAM) analysis, $q<0.001$). *CLC* had a >10-fold higher mean expression in the tumors from the early-onset group of patients as compared with that in the late-onset group, whereas the mean expression of *IFNAR1* in the early-onset group was 71% of that of the late-onset group. The gene expression of *CLC* and *IFNAR1* was validated by quantitative real-time reverse transcription-PCR and the expression differences between the early- and late-onset samples were reproduced with correlation coefficients for the gene expression values from the two analyses at 0.78 ($P<0.0001$) for *CLC* and 0.52 ($P=0.0008$) for *IFNAR1* (Figure 2).

Additional potential candidate genes, even though they were not statistically significant, were identified based on their *d*-score and fold-change values from SAM. Twenty-three genes were identified with a higher expression in the early-onset group as compared with that in the late-onset group using the following selection criteria: *d*-score or fold change >2.0, or genes with both a *d*-score and a fold change >1.5 (Table 1). Four of these genes, *CLC*, *CRI2*, *TBC1D17* and *XRCC1*, are localized within chromosome band 19q13 (including 19q13.1, 19q13.2 and 19q13.33) and three genes are encoded within the mitochondrial genome. The inverse criteria were used to select genes with a reduced expression in the early-onset group, *d*-score <−2.0 and fold change <0.5, or *d*-score <−1.5 combined with fold change <0.67, and altogether 10 genes were identified (Table 1).

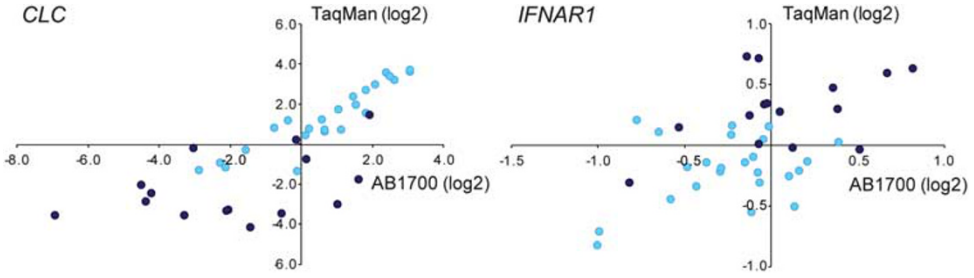

Figure 2 *CLC* and *IFNAR1* gene expression. Gene expression measurements of *CLC* and *IFNAR1* by microarrays (AB1700) and quantitative reverse transcription-PCR (TaqMan) in early-onset (light blue) and late-onset CRCs (dark blue).

Table 1 Genes with differential expression between CRCs from early- and late-onset disease identified by SAM

Rank[a]	*Probe ID*	*Gene symbol*[b]	*Gene name*	*Cytoband*[e]	*Score (d)*	*Fold change*
Higher expression in the early-onset group						
1	118354	*CLC*	Charcot–Leyden crystal protein	19q13.1	2.37	10.60
2	125930	*CRI2 (EID2)*[c]	EP300-interacting inhibitor of differentiation-2	19q13.2	1.81	1.66
3	195997	*C3AR1*	Complement component-3a receptor-1	12p13.31	1.76	2.07
4	706951	*hCG2038901*[d]		2p12	1.73	1.81
5	127594	*FGL2*	Fibrinogen-like-2	7q11.23	1.68	2.06
6	484021	*FCGR3B*|*FCGR3A*	Fc fragment of IgG, low-affinity IIIb, receptor (CD16b)|Fc fragment of IgG, low-affinity IIIa, receptor (CD16a)	1q23	1.65	1.53
7	693445	*ENSG00000198786 (MT-ND5)*[c]	Mitochondrially encoded NADH dehydrogenase-5	MT	1.61	1.96
8	220534	*AF339085*[d]		MT	1.60	1.96
9	106014	*TBC1D17*	TBC1 domain family, member-17	19q13.33	1.60	1.70
10	231270	*hCG2007748*[d]		21q21.1	1.57	1.55
11	156350	*AUTS2*	Autism susceptibility candidate-2	7q11.22	1.56	1.61
12	154454	*XRCC1*	X-ray repair-complementing-defective repair in Chinese hamster cells-1	19q13.2	1.56	1.51
13	140620	*AMICA1*	Adhesion molecule, interacts with CXADR antigen-1	11q23.3	1.53	2.20
14	145149	*MGC33657*[d]		2q14.2	1.53	1.69
15	112567	*PTAFR*	Platelet-activating factor receptor	1p35-p34.3	1.53	1.59
16	152463	*PLCL2*	Phospholipase-C-like-2	3p24.3	1.51	2.60
17	190878	*MNDA*	Myeloid cell nuclear differentiation antigen	1q22	1.51	1.98
18	336701	*ENSG00000198868 (MT-ND4L)*[c]	Mitochondrially encoded NADH dehydrogenase-4L	MT	1.51	1.60
19	199310	*FBXW4*	F-box and WD repeat domain-containing-4	10q24	1.51	1.54
20	147327	*TNFRSF25*	Tumor necrosis factor receptor superfamily, member-25	1p36.2	1.50	1.67
21	216640	*MS4A6A*	Membrane-spanning 4-domains, subfamily-A, member-6A	11q12.1	1.49	2.02
22	207163	*ITLN1*	Intelectin-1 (galactofuranose binding)	1q23.3	1.47	3.43
23	170016	*C1QA*	Complement component-1, q subcomponent, A-chain	1p36.3-p34.1	1.46	2.03
Lower expression in the early-onset group						
1	225293	*IFNAR1*	Interferon (alpha, beta and omega) receptor-1	21q22.1	−2.09	0.71
2	211400	*LOC400128*[d]		13q14.11	−1.71	0.47
3	138381	*TRAF5*	TNF receptor-associated factor-5	1q32	−1.69	0.53
4	207067	*FLJ43663*[d]		7q32.3	−1.64	0.47
5	102926	*SERPINE1*	Serpin peptidase inhibitor, clade-E (nexin, plasminogen activator inhibitor type-1), member-1	7q21.3-q22	−1.59	0.19
6	201294	*EFHC1*	EF-hand domain (C-terminal) containing-1	6p12.3	−1.56	0.59
7	120675	*SCAP1 (SKAP1)*[c]	Src kinase-associated phosphoprotein-1	17q21.32	−1.55	0.29
8	157994	*hCG2042068.1*[d]		17q21.32	−1.55	0.20
9	211045	*FAM89A*	Family with sequence similarity-89, member-A	1q42.2	−1.52	0.64
10	167664	*SNAPC1*	Small nuclear RNA-activating complex, polypeptide-1, 43 kDa	14q22	−1.52	0.64

Abbreviations: CRC, colorectal cancer; SAM, Significance Analysis of Microarrays.
Selection criteria for high expression in early-onset CRCs: SAM d-score or fold change >2, or both d-score and fold change >1.5. Selection criteria for low expression in early-onset CRCs: SAM d-score <−2.0 or fold change <0.5, or d-score <−1.5 and fold change <0.67.
[a]Genes are ranked by their order of significance.
[b]Gene symbol according to AB1700 annotation file (version, 20060930_ab1700_human).
[c]Genes having updated symbols approved by the HUGO Gene Nomenclature Committee (per 15 March 2010).
[d]An approved gene symbol from the HUGO Gene Nomenclature Committee was not available, nor was any gene identified by BLAST search.
[e]Cytogenetic band as provided by the AB1700 annotation. When information was limited to include only chromosome number, the oligo sequences and Ensembl genome information (release 57, March 2010) were used to find the specific cytogenetic band.

The expression differences of the identified genes in the samples from the early- and late-onset CRCs, and from the normal colonic mucosa samples, are presented in Figure 3. The 50 most differentially expressed genes, both on the high and on the low end for tumors from the early- as compared with the late-onset groups, are listed in Supplementary Tables 3 and 4.

In silico *analysis reveals a different immune response between early- and late-onset CRC*

The top higher and lower differentially expressed genes in the early- versus late-onset group, *CLC* and *IFNAR1*, respectively, are highly immune relevant and prompted us to examine the immune component in larger detail.[16,17] For that reason, we performed a text-mining method that profiles the global immune information content of genes from the whole of Medline.[18] Using this approach, an immune information score was assigned to the differentially expressed genes between the two groups. This resulted in the detection of a differential immune response between early- versus late-onset CRC in that there was a significant trend toward genes that had high immune relevance and increased expression in the early-onset patients as compared with that in the late-onset patients (Table 2). In addition, this *in silico* analysis showed that increased immune information content was

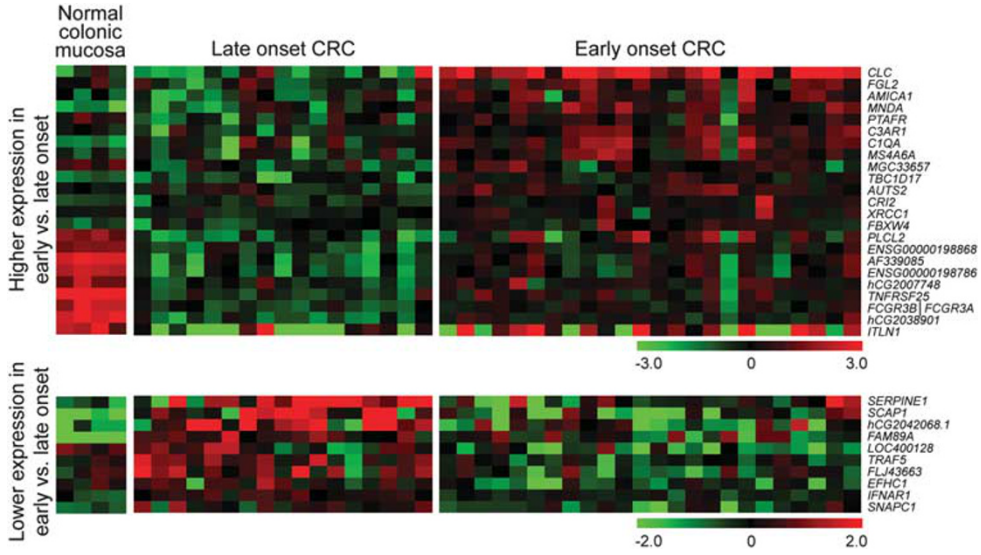

Figure 3 Differentially expressed genes between early- and late-onset CRCs. The gene expression differences of the top-scoring genes identified by SAM analysis are visualized here by a heatmap. Within each of the two sets of genes (higher and lower expressed in the early-onset CRCs), genes are clustered for improved visualization.

Table 2 Correlation of immune score and age at diagnosis and tumor stage

Clinical annotation	*Number of patients*	*Immune information score*[a]	P-*value*[b]
Age at diagnosis			
Early-onset group	24	234	0.017
Late-onset group	17	−190	0.171
Tumor stage			
Stage I	9	−510	0.016
Stage II	11	115	0.560
Stage III	15	262	0.088
Stage IV	6	297	0.278

[a]A composite immune information (from Medline) and gene expression score that quantifies the changes in the immune component of the differentially expressed genes.[18] The magnitude of positive values indicated trends toward increased expression and immune importance, and lower expression for negative values.
[b]*P*-value generated by a Monte Carlo approach (see Materials and methods) to test the significance of the immune information score.

associated with more severe tumor stages. This trend was independent of age of onset and particularly prominent for lower-expressed genes in stage I CRC (Table 2). In order to provide some understanding of the related mechanisms behind the differential immune response between the early- and late-onset groups, networks of protein–protein interactions were generated around the *CLC* and *IFNAR1* genes (Figures 4a and b). Only interaction partners that were differentially expressed between the early- and late-onset groups were allowed as partners in the networks.

Genomic regions enriched for gene expression differences between early- and late-onset CRC

The gene set enrichment analysis method was used to detect genes with coordinately higher or lower expression encoded in the same chromosome bands (Supplementary Table 5). Chromosome band 19q13, including sub-bands 19q13.1, 19q13.2 and 19q13.3, was the most statistically significant enriched region for genes with higher expression in the early-onset group. This observation was in agreement with the SAM results, in which nine out of the top 50 scoring genes were located within the 19q13 chromosome band. The genes encoded in chromosome band 13q were enriched among those with lower expression in the early-onset CRCs (Supplementary Table 5).

Identification of genes differentially expressed in a subset of the early-onset tumors

The highest-scoring genes resulting from the outlier analyses based on the 90th and 75th percentiles were partly overlapping (six out of the top 10 genes), as was also the case when comparing the results from the 25th and 10th percentiles (five out of 10 genes). By inspecting the expression plots of these top-scoring genes, it was observed that the most distinct outlier profiles in general resulted from analyses based on the 90th and 10th percentiles. The outlier expression profiles of the two highest-scoring genes/loci (*H3F3A* and *FLJ20323*; *hCG1820938.2* and *DNAJC8*) from each of these percentiles are shown in Figure 5, and the 10 genes with the highest scores from both the high and the low ends are presented in Table 3 (extended gene lists in Supplementary Tables 6 and 7). Four out of these 20 genes encode ribosomal proteins, three of which showed lower expression levels in subsets of early-onset patients and one had higher expression level.

Discussion

There is a vast amount of published data on the genetics and epigenetics of CRC,[19] but in addition to molecular changes, environmental factors contribute to tumor initiation and progression to a varying extent.[20] The

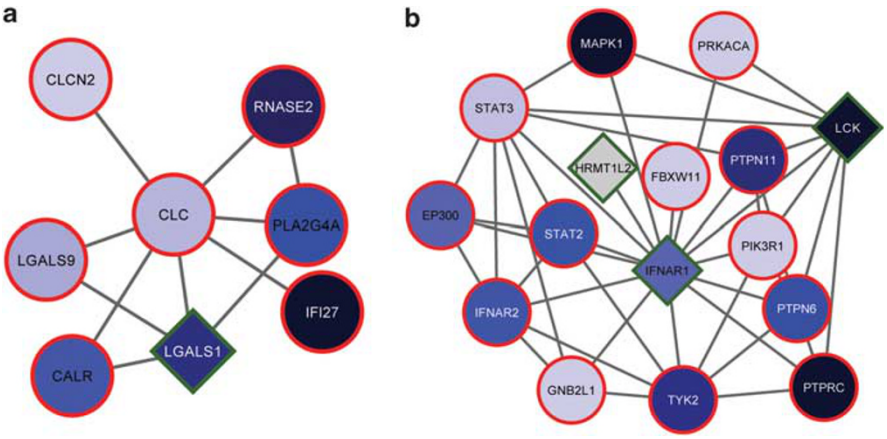

Figure 4 Protein interaction network analysis of (**a**) *CLC* and (**b**) *IFNAR1*. The genes are color-coded in blue according to their amount of immunological information in Medline.[37] Genes with a decreased expression in the comparison are shaped as diamond squares with green borders, and those with increased expression are shaped as circles with red borders. The networks have been visualized by using the Cytoscape software.[41]

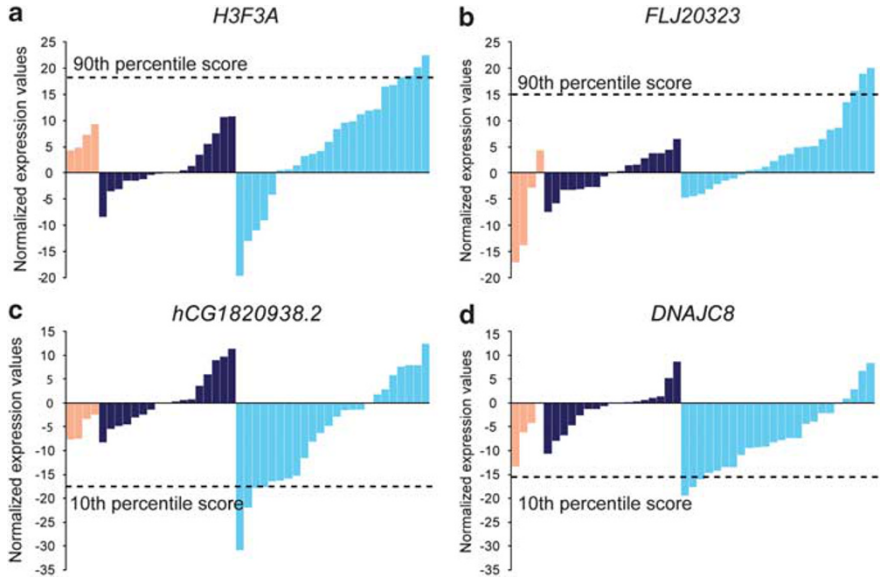

Figure 5 Genes with outlier expression profiles in early-onset CRC. The two top-scoring genes from the 90th (**a**, **b**) and 10th (**c**, **d**) percentiles are shown. The bars represent individual samples (from left: normal colonic mucosa, beige; late onset, dark blue; early onset, light blue) and are within each of sample groups sorted by the normalized expression values of their percentiles (*y*-axis). The dashed line indicates the percentile score.

distinctions between development of familial CRC with no known hereditary cancer syndrome association and sporadic CRC are not completely understood. One could expect that tumors arising at an early age are more likely to be influenced by the patient's germline genotype. In the present study, somatic gene expression differences have been identified, potentially reflecting cancer development in a predisposition context, in a group of patients clinically characterized by early-onset CRC and with no known hereditary cancer syndrome. Initial analyses of the data showed that the total sample set appears as a homogenous group, which is an advantage when searching for small, and yet undiscovered, gene expression variation in a group of patients diagnosed at an early age. Furthermore, differentially expressed genes between cancerous and normal tissue were clearly in agreement with consistently reported genes.[14,15]

To the best of our knowledge, there have been no studies published focusing on the transcriptome differences in tumors from early-onset CRC patients in order to identify novel CRC susceptibility genes. Relevant to our study, Hong *et al*.[6] have reported gene expression differences in early-onset CRC patients by comparing their normal colonic mucosa with normal colonic mucosa from healthy individuals, and identified a susceptibility gene set for early-onset CRC. Six of these seven genes were present in our data set; however, none of them were differentially expressed between the tumors from early- and late-onset patients.

Significance analysis using strict multiple testing corrections showed two genes with discriminating

Table 3 Genes with a distinct expression pattern in a subgroup of the early-onset tumors detected by outlier analysis

Rank[a]	*Probe ID*	*Gene symbol*[b]	*Gene name*	*Cytoband*[e]	*Percentile score*
90th percentile					
1	137541	*H3F3A*	H3 histone, family-3A	1q41	18.6
2	181231	*FLJ20323* (*MIOS*[c])	Missing oocyte, meiosis regulator, homolog (*Drosophila*)	7p22-p21	15.0
3	170887	*LOC285053* (*RPL18A*[f])	Ribosomal protein-L18a	19p13	12.9
4	194925	*C1orf142* (*SNAP47*)[c]	Synaptosomal-associated protein, 47 kDa	1q42.13	12.7
5	204213	*OSTF1*	Osteoclast-stimulating factor-1	9q13-q21.2	12.7
6	150916	*hCG2042834* (*ATP6V1F*)[f]	ATPase, H+ transporting, lysosomal 14 kDa, V1 subunit-F	7q32.1	12.4
7	129437	*NS4ATP2* (*SAP30L*)[c]	SAP30-like	5q33.2	12.3
8	204336	*hCG1820579.1*[d]		4q28.3	12.0
9	111502	*PHF14*	PHD finger protein-14	7p21.3	11.6
10	181269	*hCG2041316.2*[d]		1p34.1	11.6
10th percentile					
1	159501	*hCG1820938.2 (RPL7A)*[f]	Ribosomal protein-L7a	9q34	−17.8
2	142471	*DNAJC8*	DnaJ (Hsp40) homolog, subfamily-C, member-8	1p35.3	−15.7
3	219901	*ELL*	Elongation factor RNA polymerase-II	19p13.1	−15.4
4	184550	*LOC283412* (*RPL29*)[f]	Ribosomal protein-L29	3p21.3-p21.2	−12.7
5	189520	*KCMF1*	Potassium channel-modulatory factor-1	2p11.2	−12.7
6	214795	*RPS6*	Ribosomal protein-S6	9p21	−12.5
7	145714	*SNX22*	Sorting nexin-22	15q22.31	−12.0
8	175513	*MED8*	Mediator complex subunit-8	1p34.2	−11.9
9	211514	*KBTBD2*	Kelch repeat and BTB (POZ) domain-containing-2	7p14.3	−11.4
10	148299	*WDR25*	WD repeat domain-25	14q32.2	−11.2

The 10 genes with the highest (90th percentile) and lowest (10th percentile) percentile score are presented.
[a]Genes are ranked by their percentile score: 90th percentile, descending; 10th percentile, ascending.
[b]Gene symbol according to AB1700 annotation file (version, 20060930_ab1700_human).
[c]Genes having updated symbols approved by the HUGO Gene Nomenclature Committee (per 15 March 2010).
[d]An approved gene symbol from the HUGO Gene Nomenclature Committee was not available, nor was any gene identified by BLAST search.
[e]Cytogenetic band as provided by the AB1700 annotation. When information was limited to include only chromosome number, the oligo sequences and Ensembl genome information (release 57, March 2010) were used to find the specific cytogenetic band.
[f]Gene symbol, name and chromosome band were updated according to BLAST search.

expression differences between the early- and late-onset tumors, namely *CLC* and *IFNAR1*. Additional genes have also been considered as potentially interesting, although they did not reach statistical significance. *CLC* was the most differentially expressed gene, with a 10-fold higher expression in the early- versus late-onset tumors. It is known to be expressed primarily in eosinophils and basophils, and the level of CLC protein has been found to be correlated with eosinophil density in inflammation.[16] The gene shows sequence similarities with members of the galectin family and is also known as galectin-10.[21] Although the function of CLC is not known in detail, the altered expression level and distribution pattern of different galectins are suggested to act as important modulators of tumor progression in general, as well as in CRC.[22,23]

IFNAR1, showing significantly lower expression in the early- versus late-onset tumors, encodes an interferon (IFN) receptor that mediates signal transduction upon binding of type-I IFNs. Because of its effect on tumor cells, IFNs are widely used in the immunotherapy of different cancer types.[24] The IFN signaling pathway is involved in a diversity of biological processes and induces antiviral, antiproliferative and immunological responses through activation of the JAK–STAT signaling pathway.[25] Dysregulated STAT-mediated gene transcription is associated with oncogenesis, emphasizing the importance of a proper regulation of this pathway.[26] Interestingly, the present data show that *IFNAR1* has a more restricted activation in early-onset as compared with that in late-onset CRC. The expression levels for several of the genes with a lower expression in the early-onset group were similar to the expression levels seen within the normal colonic mucosa (Figure 3). It was beyond the scope of this study to illuminate whether this restricted activation in the early-onset CRCs is due to selection of different molecular pathways in the early- and late-onset groups during tumorigenesis, or whether it is caused by genetic alteration or regulatory changes.

Our detection of a global immune difference among the differentially expressed genes between early and late onset is interesting. The immune system changes during aging and becomes less effective, and, together with an accumulation of genetic and epigenetic changes, contributes to the increase of cancer incidences among elderly people. Immunity in cancer is two-sided, as inflammatory conditions can be tumor-promoting, as also shown in CRC, or the cancer activates an inflammatory response that restrains the tumor progression.[27] We have no indication of any inflammatory condition prior to CRC diagnosis leading to an elevated cancer risk within the early-onset sample group, although this cannot be excluded. If the increased immune content is associated to age, this supports the hypothesis of the existence of underlying genetic alteration causing an early onset of disease. There is increasing evidence of immune signatures as prognostic factors in CRC,[28] and a consequence of our observation could be a better prognosis among the early-onset patients. However, our data showed an association between an increase of genes

associated with immune response and a more severe tumor stage.

In addition to *CLC*, *CRI2*, *TBC1D17* and *XRCC1* also showed higher expression in the early-onset group than in the late-onset group (Table 1), as well as in comparison with the normal colonic mucosa (Figure 3). These are all located within the same chromosome band, 19q13. Interestingly, the sub-band 19q13.1 has by a genome-wide association study been highlighted as potentially carrying a susceptibility locus for CRC.[29] Although the individual top-scoring genes highlighted in this study are megabases away from this identified susceptibility locus, chromosomal sub-bands at 19q13 were identified among the most significantly enriched regions for differentially expressed genes in early- versus late-onset CRC. The enriched regions can potentially indicate genomic copy-number changes or long-range regulatory mechanisms. Array comparative genomic hybridization data available from the same patients as analyzed in the present study confirmed that the locus containing *CLC* has frequent DNA copy-number gain in CRC, and preferentially in early-onset tumors.[8]

After *CLC*, *ITLN1* had the second highest fold change when comparing early- and late-onset tumors. Interestingly, a recent study identified variants of *ITLN1* as susceptibility loci to Crohn's disease.[30] Compared with the normal colonic mucosa, a higher expression of *ITLN1* was in general seen in tumors from the early-onset group, whereas the late-onset group showed a lower *ITLN1* expression.

A challenge in the search for predisposing genetic changes is the heterogeneity within the group of early-onset cancers. This may be due to both the existence of different predisposition types, and also due to inclusion of patients with sporadic disease, who accidentally have developed cancer at a young age. Outlier statistics has the potential to identify subgroups within a defined group. Here, we used this statistical approach to identify gene expression changes, which are only present within a few early-onset CRC samples. This analysis revealed transcriptional differences of several genes, including *H3F3A*, a member of the histone H3 family that, in association with other histones, is involved in transcriptional regulation. Further, four ribosomal protein-coding genes (*RPL18A*, *RPL7A*, *RPL29* and *RPS6*) were found differentially expressed within subgroups of early-onset CRC. As ribosomes are the main players in protein synthesis, they are consequently important regulators of cell proliferation and growth. The defects in ribosome biogenesis have been shown to increase susceptibility to cancer in several inherited genetic disorders.[31]

In conclusion, several genes have been identified with expression differences between cancer tissues obtained from early- and late-onset CRC patients. These two groups were similar at the overall gene expression level, providing additional support to the importance of the discriminated genes in the development of CRC within a genetic susceptibility context.

Materials and methods

Patients and tumor samples

The CRC patients included in the present study had all undergone primary surgery at hospitals located in the South East region of Norway. Twenty-four patients were diagnosed at a young age (mean, 43 years; range, 28–53 years), referred to as the early-onset group. They were excluded from the HNPCC and familial adenomatous polyposis syndromes by clinical criteria and no other cancer syndromes were recorded for these patients. The second group consisted of 17 patients with primary diagnosis at old age (mean, 79 years; range, 69–87 years), referred to as the late-onset group. The samples from the late-onset group were selected to reflect the composition of the early-onset group with respect to gender, tumor localization and tumor stage according to The International Union Against Cancer/American Joint Committee on Cancer. Microsatellite instability status was analyzed previously in all samples to eliminate the potential risk of including patients with inherited DNA-repair deficiencies.[8] A summary of clinical data related to the two groups of early- and late-onset CRC is provided in Supplementary Table 1. Normal colonic mucosa was taken from disease-free areas distant to the primary tumors of four CRC patients from the late-onset group (mean, 78 years; range, 66–83 years). Two of these samples had the corresponding primary tumor analyzed.

Tumor tissues and normal colonic mucosa from the late-onset group were snap-frozen in liquid nitrogen before storage at −80 °C. Samples from the early-onset group were transferred to the RNAlater RNA Stabilization Reagent (Qiagen, Hilden, Germany), followed by removal of the RNAlater liquid and long-term storage at −80 °C. In addition, a tissue section in close proximity to the specimen used for RNA extraction was evaluated by a pathologist for tumor cell content and quality of the tissue. This procedure was limited to include tissue from the late-onset patients. Owing to change in tissue properties after storage in RNAlater, histological re-evaluation was not possible for the samples from the early-onset group.

RNA from fresh frozen tumor tissue samples was extracted by using the AllPrep DNA/RNA Mini Kit (Qiagen) and RNA from normal colonic mucosa was isolated by using the Ambion RiboPure kit (Applied Biosystems, Foster City, CA, USA) according to the manufacturers' protocols. RNA was quantified by UV spectroscopy (NanoDrop ND-1000; Thermo Fisher Scientific, Waltham, MA, USA) and quality was assessed by using the Agilent 2100 Bioanalyzer (Agilent Technologies, Santa Clara, CA, USA). All RNA samples included in the downstream analyses had RNA integrity values above 8.

Written informed consent was obtained from all subjects included. The research biobanks have been registered according to national legislation and the study has been approved by the Regional Committee for Medical Research Ethics (REK South-East: 1.2005.1629; REK South: 2003, S-02126).

Gene expression microarray procedure

Tumor samples were randomized prior to gene expression analyses. Gene expression analysis was performed by using the Applied Biosystems 1700 microarray platform and the Human Genome Survey Microarray V2.0 containing 32 878 unique 60-mer oligonucleotide probes (Applied Biosystems). A 1-μg of total RNA was amplified and converted to digoxigenin-labeled cRNA by using the NanoAmp RT-IVT Labeling kit (Applied Biosystems) in

accordance with the manufacturer's protocol. The Chemiluminescent Detection kit (Applied Biosystems) was used for the preparation of labeled cRNA, and hybridization and washing solutions. Both digoxigenin-UTP and the anti-digoxigenin -AP were provided by Roche Diagnostics (Mannheim, Germany). Digoxigenin-labeled probes were hybridized to the microarrays overnight (16 h) at 55 °C and chemiluminescence signals were detected by using AB1700 Chemiluminescent Microarray Analyzer (Applied Biosystems). Image preparation, signal intensity quantification and initial analysis were performed by using the accompanying software from Applied Biosystems (version 1.1.1) using default settings.

Analysis of microarray data

The Bioconductor package ABarray (Applied Biosystems) was used for quality assessment and to generate quantile-normalized data (http://bioconductor.org/packages/1.9/bioc/html/ABarray.html). Only probes for which the normalized signals were more than three standard deviations greater than the local background levels, in at least half of the samples, were included in the downstream data analyses. Values with low quality scores (flag code >8191) were noted as missing, as recommended by the vendor. The remaining missing values were imputed by the K-nearest neighbor algorithm ($k=10$) by using the Microsoft Excel add-in Significance Analysis of Microarrays (SAM, version 3.09; http://stat.stanford.edu/~tibs/SAM/). Genes that initially had flagged probes in more than 10% of the samples were removed and not applied to the downstream analyses. For each probe, the gene expression values were centered over the median across all the samples. A batch variation was seen in the data set and by using SAM analysis the 320 most affected probes were identified and removed from further analyses. The resulting data set included 17 085 probes. The features were annotated according to the latest information provided by the manufacturer ('20060930_ab1700_human.txt'; downloaded from the Panther website; http://www.pantherdb.org/). Genes presented in this paper, which were not annotated by gene name or function in the microarray annotation file, were submitted to the Basic Local Alignment Search Tool (BLAST) from The National Centre for Biotechnology Information by using the corresponding oligo sequences and RNA reference sequences as target database (http://blast.ncbi.nlm.nih.gov/Blast.cgi).

Raw data are publicly available in the GEO database (accession number GSE25071).

Sample characteristics. To illustrate the relationship between individual samples, principal component analysis and hierarchical cluster analysis (Euclidean distance metrics; average linkage clustering) were performed with inclusion of gene expression values from the 1350 probes with the largest ranges between the lowest and the highest expression values. Both analyses were performed in J-express (J-express pro, version 2.9; MolMine, Bergen, Norway).

Differences between cancer and normal samples. Differentially expressed genes between CRC and normal colonic mucosa were identified by SAM (*t*-statistics on median-centered probe values). To illustrate expression differences in tumor and normal colonic mucosa, a heatmap was generated by using MultiExperiment Viewer (MeV, version 4.5.1; http://www.tm4.org/mev).

Differences between early- and late-onset cancer. Two approaches were used to identify gene expression differences between the two groups of patients. First, SAM analysis, using Wilcoxon statistics on median-centered probe values, were performed to identify genes whose expression was generally different between the two sample groups. Second, an outlier analysis was used to detect gene expression differences, which are present in only a fraction of the samples in the early-onset group. The statistical approach was in essence similar to the one described by Tomlins *et al*.[32] Here, the late-onset samples were used as a reference group, and for each gene, the median $\log_2$ value and the variance across these samples were calculated. Subsequently, expression values from the normal colonic mucosa samples and the early- and late-onset samples were centered over this median value and variation was standardized by dividing by the variance in the late-onset samples. As measures for outlier expression, the 90th, 75th, 25th and 10th percentiles from the early-onset group were calculated. Finally, we excluded genes for which any of the normal samples had values exceeding the percentile scores.

Immune scoring of cancer gene expression. To assign immune associations to the differentially expressed genes, we used a recent method that profiles the global immune information content of genes from the whole of Medline. The details of this method are outlined in Clancy *et al*.[18] Briefly, it integrates manual curation of relevant immune terms, text mining of the Medline database and a scoring procedure that applies information theory to assign an immune information score to every human gene. The immune information score was then used to create a composite immune and gene expression value. In this process, each probe signal intensity measurement was assigned a fold change relative to that probe's mean signal intensity across all samples and used to create a weighted composite signal intensity and immune information score for each gene. This weighted composite score for each gene was then summated across all genes for each patient to generate a weighted immune score for each patient. These scores were then compared to the clinical annotations (age at onset and stage) to find correlations between the weighted immune score and the clinical phenotypes. For statistical analysis of these comparisons, Monte Carlo simulations with 10 000 draws were used to create a null distribution for each comparison. Pearson's correlation was used for phenotypes with a numerical descriptive.

Protein interaction network analysis. A protein interaction network was generated around IFNAR1 from an integrated set of three different protein interaction databases.[33–35] For the CLC network, the highest-ranked interactions were retrieved from a resource that integrates the protein interaction phrases from the Biomolecular Interaction Network Database and their co-citations with gene names in the same sentence from the over 20 million abstracts in Medline.[36,37] A gene has a connection to a neighboring gene in both networks only if the

neighboring gene was differentially expressed between the early- versus late-onset groups.

Gene set enrichment analysis. Gene set enrichment analysis was performed in R (version 2.9.2) by using the Category package (version 2.5) in Bioconductor (version 2.5).[38] This analysis allows the identification of consistent differential expression of genes mapped to chromosomal bands as defined in the UCSC database.[39] Expression data, pre-processed as described previously, were used and only features annotated with Entrez gene identifiers were included. An empirical Bayesian statistic was used to identify differentially expressed genes between early- and late-onset patients, followed by fitting the gene set enrichment analysis model for finding the enriched chromosomal bands.[40] A significance threshold was set at $P<0.001$.

Validation of gene expression data

Two genes, *CLC* and *IFNAR1*, were selected for technical validation by quantitative real-time reverse transcription-PCR, performed by using TaqMan 7900 HT (Applied Biosystems). Both genes were significantly differentially expressed in early- versus late-onset tumors. A 1-μg weight of total RNA was reverse-transcribed into cDNA by using the High Capacity RNA-to-cDNA kit (Applied Biosystems) as described by the manufacturer. TaqMan gene expression assays were pre-designed by Applied Biosystems (*CLC*, hs00171342_m1; *IFNAR1*, hs00265057_m1). Real-time PCRs were performed in triplicate, using 10 ng of cDNA in each reaction. Both assays were performed by using the TaqMan Fast Universal PCR Mastermix (No AmpErase UNG; Applied Biosystems) in a total volume of 10 μl and under recommended thermal cycling conditions (pre-activation at 95 °C for 15 s followed by 40 cycles of 95 °C for 1 s, and 60 °C for 20 s). A standard curve was prepared by serial dilution of cDNA from the human universal reference RNA (UHR; Stratagene, La Jolla, CA, USA). Two endogenous controls were analyzed in parallel with each assay (*GUSB*, 4333767F and *ACTB*, 4352935E; Applied Biosystems). To calculate the quantity of *IFNAR1*, relative quantification was used. The median from the two endogenous controls was used to normalize the expression values in each sample. *CLC* was only weakly expressed in the UHR, and consequently, expression values were calculated by using the comparative C_T method. Compliance between AB1700 and TaqMan data was assessed by Pearson's correlation (SPSS software, version 17.0).

Conflict of interest

The authors declare no conflict of interest.

Acknowledgements

This study was funded by grants from the Norwegian Cancer Society to GEL, RAL (including PhD grants to THÅ and MB) and RIS; by a grant from the South-Eastern Norway Regional Health Authority to RAL (including a postdoctoral grant to LC); and grants to ETE from the Norwegian Foundation for Health and Rehabilitation through the EXTRA funds. We thank all members of the INFAC study group, Kari Almendingen, Arne Bakka, Gunter Bock, Torunn Fetveit, Hans Joachim Hauss, Anders Husby, Ragnhild A Lothe, Tom Mala, Øystein Mathisen, Ingvild Moberg, Arild Nesbakken, Arve Rennesund, Oddvar Sandvik, Espen Thiis-Evensen and Morten H Vatn, for providing us with tissue samples and clinical data from the CRC patients diagnosed at a young age.

References

1 Altekruse SF, Kosary CL, Krapcho M, Neyman N, Aminou R, Waldron W *et al.* SEER Cancer Statistics Review, 1975–2007. *National Cancer Institute* (http://www.seer.cancer.gov/csr/1975_2007/browse_csr.php) 2010.
2 Rustgi AK. The genetics of hereditary colon cancer. *Genes Dev* 2007; **21**: 2525–2538.
3 de la Chapelle A. Genetic predisposition to colorectal cancer. *Nat Rev Cancer* 2004; **4**: 769–780.
4 Skoglund J, Djureinovic T, Zhou XL, Vandrovcova J, Renkonen E, Iselius L *et al.* Linkage analysis in a large Swedish family supports the presence of a susceptibility locus for adenoma and colorectal cancer on chromosome 9q22.32-31.1. *J Med Genet* 2006; **43**: e7.
5 Blaker H, Mechtersheimer G, Sutter C, Hertkorn C, Kern MA, Rieker RJ *et al.* Recurrent deletions at 6q in early age of onset non-HNPCC- and non-FAP-associated intestinal carcinomas. Evidence for a novel cancer susceptibility locus at 6q14-q22. *Genes Chromosomes Cancer* 2008; **47**: 159–164.
6 Hong Y, Ho KS, Eu KW, Cheah PY. A susceptibility gene set for early onset colorectal cancer that integrates diverse signaling pathways: implication for tumorigenesis. *Clin Cancer Res* 2007; **13**: 1107–1114.
7 Tenesa A, Dunlop MG. New insights into the aetiology of colorectal cancer from genome-wide association studies. *Nat Rev Genet* 2009; **10**: 353–358.
8 Berg M, Agesen TH, Thiis-Evensen E, Infac IS, Merok MA, Teixeira MR *et al.* Distinct high resolution genome profiles of early onset and late onset colorectal cancer integrated with gene expression data identify candidate susceptibility loci. *Mol Cancer* 2010; **9**: 100.
9 Koessler T, Oestergaard MZ, Song H, Tyrer J, Perkins B, Dunning AM *et al.* Common variants in mismatch repair genes and risk of colorectal cancer. *Gut* 2008; **57**: 1097–1101.
10 Giraldez MD, Balaguer F, Caldes T, Sanchez-de-Abajo A, Gomez-Fernandez N, Ruiz-Ponte C *et al.* Association of MUTYH and MSH6 germline mutations in colorectal cancer patients. *Fam Cancer* 2009; **8**: 525–531.
11 Quackenbush J. Microarray analysis and tumor classification. *N Engl J Med* 2006; **354**: 2463–2472.
12 Perou CM, Sorlie T, Eisen MB, van de RM, Jeffrey SS, Rees CA *et al.* Molecular portraits of human breast tumours. *Nature* 2000; **406**: 747–752.
13 Sorlie T, Perou CM, Tibshirani R, Aas T, Geisler S, Johnsen H *et al.* Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proc Natl Acad Sci USA* 2001; **98**: 10869–10874.
14 Cardoso J, Boer J, Morreau H, Fodde R. Expression and genomic profiling of colorectal cancer. *Biochim Biophys Acta* 2007; **1775**: 103–137.
15 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**: 543–552.
16 De Re V, Simula MP, Cannizzaro R, Pavan A, De Zorzi MA, Toffoli G *et al.* Galectin-10, eosinophils, and celiac disease. *Ann NY Acad Sci* 2009; **1173**: 357–364.

17 Bracarda S, Eggermont AM, Samuelsson J. Redefining the role of interferon in the treatment of malignant diseases. *Eur J Cancer* 2010; **46**: 284–297.
18 Clancy T, Pedicini M, Castiglione F, Santoni D, Nygaard V, Lavelle TJ *et al*. Immunological network signatures of cancer progression and survival. *BMC Med Genomics* 2011; **4**: 28.
19 Grady WM, Carethers JM. Genomic and epigenetic instability in colorectal cancer pathogenesis. *Gastroenterology* 2008; **135**: 1079–1099.
20 Lichtenstein P, Holm NV, Verkasalo PK, Iliadou A, Kaprio J, Koskenvuo M *et al*. Environmental and heritable factors in the causation of cancer—analyses of cohorts of twins from Sweden, Denmark, and Finland. *N Engl J Med* 2000; **343**: 78–85.
21 Dyer KD, Handen JS, Rosenberg HF. The genomic structure of the human Charcot–Leyden crystal protein gene is analogous to those of the galectin genes. *Genomics* 1997; **40**: 217–221.
22 Liu FT, Rabinovich GA. Galectins as modulators of tumour progression. *Nat Rev Cancer* 2005; **5**: 29–41.
23 Demetter P, Nagy N, Martin B, Mathieu A, Dumont P, Decaestecker C *et al*. The galectin family and digestive disease. *J Pathol* 2008; **215**: 1–12.
24 Bracarda S, Eggermont AM, Samuelsson J. Redefining the role of interferon in the treatment of malignant diseases. *Eur J Cancer* 2010; **46**: 284–297.
25 Billiau A. Interferon: the pathways of discovery I. Molecular and cellular aspects. *Cytokine Growth Factor Rev* 2006; **17**: 381–409.
26 Spano JP, Milano G, Rixe C, Fagard R. JAK/STAT signalling pathway in colorectal cancer: a new biological target with therapeutic implications. *Eur J Cancer* 2006; **42**: 2668–2670.
27 Ferrone C, Dranoff G. Dual roles for immunity in gastrointestinal cancers. *J Clin Oncol* 2010; **28**: 4045–4051.
28 Galon J, Costes A, Sanchez-Cabo F, Kirilovsky A, Mlecnik B, Lagorce-Pages C *et al*. Type, density, and location of immune cells within human colorectal tumors predict clinical outcome. *Science* 2006; **313**: 1960–1964.
29 Houlston RS, Webb E, Broderick P, Pittman AM, Di Bernardo MC, Lubbe S *et al*. Meta-analysis of genome-wide association data identifies four new susceptibility loci for colorectal cancer. *Nat Genet* 2008; **40**: 1426–1435.
30 Barrett JC, Hansoul S, Nicolae DL, Cho JH, Duerr RH, Rioux JD *et al*. Genome-wide association defines more than 30 distinct susceptibility loci for Crohn's disease. *Nat Genet* 2008; **40**: 955–962.
31 Montanaro L, Trere D, Derenzini M. Nucleolus, ribosomes, and cancer. *Am J Pathol* 2008; **173**: 301–310.
32 Tomlins SA, Rhodes DR, Perner S, Dhanasekaran SM, Mehra R, Sun XW *et al*. Recurrent fusion of TMPRSS2 and ETS transcription factor genes in prostate cancer. *Science* 2005; **310**: 644–648.
33 Kerrien S, am-Faruque Y, Aranda B, Bancarz I, Bridge A, Derow C *et al*. IntAct—open source resource for molecular interaction data. *Nucleic Acids Res* 2007; **35**: D561–D565.
34 Keshava Prasad TS, Goel R, Kandasamy K, Keerthikumar S, Kumar S, Mathivanan S *et al*. Human protein reference database—2009 update. *Nucleic Acids Res* 2009; **37**: D767–D772.
35 Stark C, Breitkreutz BJ, Reguly T, Boucher L, Breitkreutz A, Tyers M. BioGRID: a general repository for interaction datasets. *Nucleic Acids Res* 2006; **34**: D535–D539.
36 Bader GD, Betel D, Hogue CW. BIND: the Biomolecular Interaction Network Database. *Nucleic Acids Res* 2003; **31**: 248–250.
37 Jenssen TK, Laegreid A, Komorowski J, Hovig E. A literature network of human genes for high-throughput analysis of gene expression. *Nat Genet* 2001; **28**: 21–28.
38 Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA *et al*. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci USA* 2005; **102**: 15545–15550.
39 Karolchik D, Kuhn RM, Baertsch R, Barber GP, Clawson H, Diekhans M *et al*. The UCSC genome browser database: 2008 update. *Nucleic Acids Res* 2008; **36**: D773–D779.
40 Sarkar D, Falcon S, Gentleman R. Using categories defined by chromosome bands. Bioconductor, version 2.5, http://www.bioconductor.org, 2008.
41 Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D *et al*. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res* 2003; **13**: 2498–2504.

Supplementary Information accompanies the paper on Genes and Immunity website (http://www.nature.com/gene)