Immunogenetics (2005) 57: 709–716
DOI 10.1007/s00251-005-0044-7

ORIGINAL PAPER

Austin L. Hughes

Consistent across-tissue signatures of differential gene expression in Crohn's disease

Received: 19 May 2005 / Accepted: 10 August 2005 / Published online: 28 September 2005

Abstract An approach based on analysis of variance was applied to raw expression data on 44,760 transcripts in order to identify those with significant differential expression across ileum and colon in Crohn's disease (CD) and ulcerative colitis (UC). The design treated tissue as a block effect, thereby removing this effect statistically and increasing the power to test for effects of disease states (control, CD, and UC). A significant *F*-statistic for the disease effect was not correlated with the ratios CD/control or UC/control, evidently because many transcripts with high-expression ratios to the control showed inconsistent patterns across tissues. Of 1,053 transcripts showing a significant effect of disease state at the 1% level by the bootstrap test, 508 showed significant difference at the 1% level in a post hoc test for difference between the mean scores for CD and control. These included a number of genes relevant to the mechanism of pathogenesis of CD and a number of genes mapping to genomic regions that have previously shown linkage to CD in association studies.

Keywords Crohn's disease · Gene expression · Inflammatory bowel disease · Microarray data

Introduction

Inflammatory bowel disease (IBD), including Crohn's disease (CD) and ulcerative colitis (UC), is characterized by chronic inflammation of the intestine in the absence of an obvious pathogenic cause, and the underlying disease mechanisms remain poorly understood. There is evidence that both genetic and environmental factors play a role in the etiology of IBD (Watts and Satsangi 2002; Bouma and Strober 2003), and that IBD may be a complex of diseases with different etiologies (Gasche et al. 2003). A significant development in recent years has been the discovery of an association between certain polymorphisms at the CARD15/NOD2 locus on chromosome 16 and increased susceptibility to CD (Hugot et al. 2001). This locus, which maps to chromosome 16q12, enodes a protein (CARD15) that uses leucine-rich repeats (LRR) to bind bacterial peptidoglycan and subsequently is involved in the activation of NF-κB Russell et al. 2004). There is evidence of at least six other susceptibility loci for IBD, including one on chromosome 12 (mapped to 12p13.2–q24.1), one on chromosome 19 (mapped to 19p13), one on chromosome 1 (1p36), one on chromosome 5 (5q31), and one on chromosome 14 (mapped to 14q11–q12), as well as the *HLA* region on chromosome 6 (Cho et al. 2000; Watts and Satsangi 2002; Girardin et al. 2003; van Heel et al. 2005; Negoro et al. 2005).

The analysis of gene expression by techniques such as microarray holds promise for increasing our understanding of both the causes and the pathology of complex diseases such as IBD (Devauchelle et Chiocchia 2004; Dieckgraefe et al. 2000; Heller et al. 1997; Kok et al. 2004; Langmann et al. 2004; Mannick et al. 2004). However, gene expression data pose difficult problems of interpretation and analysis. First of all, gene expression itself is a complex phenomenon, with potential variation arising not only from differences among tissue types and disease states but also from individual genetic differences and environmental effects. In addition, because of the cost of gene expression experiments, a typical microarray data set contains information on the expression levels of numerous transcripts, but usually, the number of replicates is small. Moreover, certain highly expressed transcripts show the most marked expression level differences between disease and normal tissues. Yet expression levels of these highly expressed transcripts may be subject to substantial stochastic error,

Electronic supplementary material Supplementary material is available for this article at http://dx.doi.org/10.1007/s00251-005-0044-7 and accessible for authorised users.

A. L. Hughes (✉)
Department of Biological Sciences,
University of South Carolina,
Coker Life Sciences Bldg., 700 Sumter St.,
Columbia, SC 29208, USA
e-mail: austin@biol.sc.edu
Tel.: +1-803-7779186
Fax: +1-803-7774002

and thus, the observed differences may not be biologically significant.

One approach to overcoming these problems in microarray data interpretation is to make use of comparisons among different tissues as well as among different states of disease. Using analysis of variance, it is possible to test for differences among disease states controlling statistically for the difference among tissues. Such an approach can be used to detect transcripts which are consistently increased or decreased in a given disease state across tissues. The identification of transcripts showing a consistent pattern across tissues serves to minimize the effects of stochastic variations in the expression of highly expressed transcripts in a given experiment.

Here I apply this approach to analyze data on gene expression in IBD from a published study that focused on dysregulation of pregnane X receptor target genes (Langmann et al. 2004). The data are raw expression scores for both ileum and colon in controls, CD patients, and UC patients. Note that, because UC is a disease of the colon, it was not expected that there would be many transcripts with significant differential expression across both ileum and colon in UC. Nonetheless, the inclusion of data from UC has the desirable property of increasing the power of the statistical analysis, by providing what amounts to an additional control and by increasing the error degrees of freedom for the analysis of variance.

Methods

Raw expression data from microarray experiments were downloaded from the Gene Expression Omnibus (GEO) database (Barrett et al. 2005). A given data set in the GEO database (a GDS record) represents a collection of biologically and statistically comparable samples. Two data sets were used: GDS559, derived from Affymetrix (Santa Clara, CA) GeneChip Human Genome U133 Array Set HG-U133Al and GDS560, derived from Affymetrix GeneChip Human Genome U133 Array Set HG-U133B. These chips provide a broad coverage of transcripts from the human genome. Each set contained measurements for two tissues, terminal ileum and colon transversum, from unaffected controls, from patients with CD, and from patients with UC. For each of the six combinations of tissue and disease state, tissue was obtained by pooling tissue from four donors. GDS559 provided data for 22,283 transcripts, and GDS560 provided data for 22,645 transcripts. Only 168 transcripts were in common between the two data sets. By examination of functional annotations, these 168 transcripts did not appear to be atypical of the data set as a whole. In the case of these 168 transcripts, I averaged the scores for these two data sets. Thus, the final data matrix contained measurements for 44,760 transcripts, providing extensive coverage of well-substantiated human genes.

The 168 transcripts shared between the data sets provided a test for the comparability of the results in the two data sets. For the six combinations of tissue and disease state, the correlations between the raw scores for these 168 transcripts in the two data sets ranged between 0.953 and 0.987 ($P<0.001$ in all cases). This result supports the hypothesis that experimental conditions in the two data sets were comparable.

For each transcript, analysis of variance was conducted in a block design. The tissue (ileum or colon) constituted the block effect (Supplementary Table S1). I tested for differences between disease states (control, CD, and UC) after removing the effect of difference among tissues. A randomization procedure was used to provide probability levels for F-statistics. Data vectors were generated for 1,000,000 simulated transcripts by sampling (with replacement) from each column of the original data matrix. The F-statistic was then calculated for each simulated transcript, and the distribution of the F-statistics for the simulated transcripts was used as a reference to assess significance of F-statistics computed from the real data. Each F-statistic computed from the real transcripts was considered significant at the α level if 100α% or fewer of the simulated transcripts showed F-statistics greater than that value. For transcripts showing a significant F-statistic, post hoc comparisons among individual disease state means were conducted by Tukey's honestly significant difference (HSD) method (Sokal and Rohlf 1981). Significance for HSD was also assessed by comparison with those calculated for the simulated transcripts. In order to correct for multiple testing, I applied the step-up false-discovery rate (FDR) method of Benjamini and Hochberg (1995) to both F tests and HSD.

Results

When analysis of variance was applied to expression data for 44,760 transcripts in the two tissues (ileum and colon) and three disease states (control, CD, and UC), there were striking differences among transcripts with respect to the proportion of the total sum of squares accounted for by differences among tissues and the proportion of the total sum of squares accounted for by disease state (Fig. 1). Almost every possible combination of values was seen (Fig. 1). There were transcripts for which disease state accounted for a very high proportion (nearly 100%) of the total sum of squares and tissue accounted for very little of the total sum of squares (Fig. 1). Conversely, there were transcripts for which disease state accounted for very little of the total sum of squares, whereas tissue accounted for a high proportion (Fig. 1). A group of 5,046 transcripts (11.3% of total) showed significant effects of disease state at the 5% level by the F test and a FDR of less than 5%. A group of 1,053 transcripts (2.4%) showed significant effects of disease state at the 1% level and FDR of less than 1%.

In the analysis of variance conducted here, it was not possible to test for interactions of tissue and disease state because of the lack of replication. In order to assess the possible impact of replication on these data, an analysis of variance testing for main effects (tissue and disease state) plus their interaction was applied to the 168 transcripts

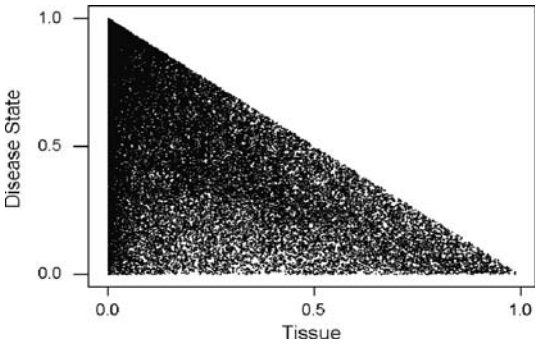

Fig. 1 Scatterplot of the percentage of the total sum of squares accounted for by disease state vs the percentage accounted for by tissue in analyses of variance for 44,760 transcripts

which were replicated in the GDS559 and GDS560 data sets (see Methods). None showed a significant tissue-by-disease-state interaction at the 5% level. Yet when the values for the two data sets were averaged to provide overall scores for these 168 transcripts, 15 of 168 (9.8%) showed a significant effect of disease state at the 5% level, and 3 (1.8%) showed a significant effect of disease state at the 1% level. These percentages are similar to those for the complete data set of 44,760 transcripts.

Transcripts showing a high value of the *F*-statistic for disease state did not necessarily show unusually high ratios of the scores for CD or UC to the control. When the *F*-statistic for disease state was plotted against the natural logarithm of the mean ratio CD/control, there was no correlation (r=0.006; ns; Fig. 2a). Rather, the transcripts with both the highest and lowest log ratios showed very low *F*-statistics, whereas the transcripts with high *F*-statistics tended to have log ratios close to zero (Fig. 2a). In the case of UC, there was a small but significant negative correlation between the *F*-statistic for disease state and the natural logarithm of the mean ratio UC/control (r=−0.016′; P=0.001; Fig. 2b). Here also, transcripts with both very high and very low log ratios tended to have low *F*-statistics, while those with high *F*-statistics tended to have log ratios close to zero (Fig. 2b). On the other hand, the natural logarithm of the mean ratio CD/control was highly, positively correlated with the natural logarithm of the ratio UC/control (r=0.559; P<0.001; Fig. 3).

A similar pattern was seen when the logarithm of maximum ratio for the two tissues of the score for CD to that of controls was correlated with the *F*-statistic; there was no significant linear relationship (r=−0.007; ns). In the case of UC, there was a small but significant negative correlation (r=−0.031; P<0.001) between the logarithm of maximum ratio for the two tissues and the *F*-statistic. These results show that neither mean nor maximum of the ratio of the scores for either CD or UC to those for controls was a good predictor of the magnitude of the effect of disease state in the analysis of variance.

Within the group of 5,046 transcripts showing a significant effect of disease state at the 5% level and FDR less than 5%, there was a group of 1,647 transcripts showing a significant HSD at the 5% level between means for CD and control and FRD less than 5%. By contrast, there were only 63 transcripts showing a significant HSD at the 5% level between means for UC and control and FDR of less than 5%. Within the group of 1,053 transcripts show-

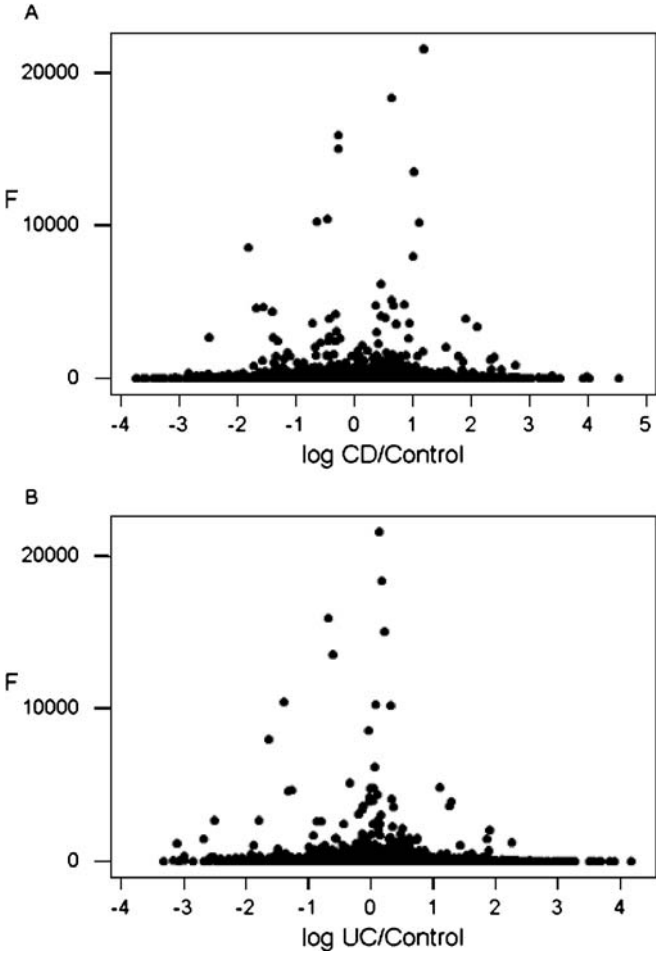

Fig. 2 Plots of the *F*-statistic for the effect of disease state vs the natural logarithm of the mean ratios. CD to control (**a**) (r=0.006; ns) and UC to control (**b**) (r=−0.016; P=0.001)

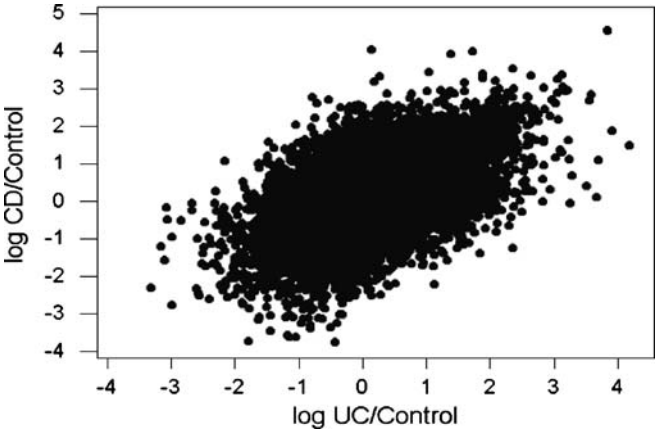

Fig. 3 Plot of the ratio of the natural logarithm of the mean ratio of CD to control vs that of the mean ratio of UC to control (r=0.559; P<0.001)

Table 1 Transcripts with annotated function from RefSeq database showing a significant effect of disease state and significant HSD for CD vs control (both at 1% level)[a]

No.	Accession number	Protein function	Map location	Notes[b]	CD[c]
1	NM_030918	Sorting nexin family member 27	1q21.3		↑
2	NM_022365	DNAJ (Hsp40) homology, subfamily C, member 1	10p12.31		↓
3	NM_014852	RNA-binding protein	12q24.21		↓
4	NM_001433	Endoplasmic reticulum to nucleus 1, transcript variant 1	17q24.2		↑
5	NM_004532	Mucin 4, transcript variant 4	3q29		↓
6	NM_016628	WW domain-containing adoptor with coiled coil, transcript variant 1	10p12.1		↑
7	NM_013390	Transmembrane protein 2	9q13–q21		↑
8	NM_005875	Translation factor su1 homolog (GC20)	3p22.1		↑
9	NM_004230	Endothelial differentiation, sphingolipid G-protein-coupled receptor 5 (EDG5)	19p13.2	L	↓
10	NM_006767	Leucine-zipper-like transcription regulator 1[b]	22q11.21	TF	↓
11	NM_012248	Selenophosphate synthase 2	16p11.2		↑
12	NM_007169	Phosphatidyl ethanolamine *N*-methyltransferase, transcript variant 2	17p11.2		↓
13	NM_006193	Paired box gene 4 (*PAX4*)	7q32	TF	↓
14	NM_001310	cAMP-responsive element binding protein-like 2 (CREBL2)	12p13	L, TF	↑
15	NM_002906	Radixin	11q23		↓
16	NM_012290	Tousled-like kinase	2q31.1		↑
17	NM_001731	B cell tranlocation gene 1, anti-proliferative (*BTG1*)	12q22	L, Im	↑
18	NM_003327	Tumor necrosis factor receptor superfamily, member 4 (*TNFRSF4*)	1p36	L, Im	↑
19	NM_025168	Leucine-rich repeat-containing 1	6p12.1		↓
20	NM_014857	RAB GTPase activation protein-like 1	1q24		↑
21	NM_000072	CD36, transcript variant 3	7q11.2		↑
22	NM_005877	Splicing factor 3a, subunit 1	22q12.2		↓
23	NM_005238	ETS1 oncogene	11q23.3	TF	↓
24	NM_004605	Sulfotransferase family, cytosolic 2B, member 1, transcript variant 1	19q13.3		↓
25	NM_018095	Kelch repeat and BTB domain-containing 4, transcript variant 1	11p11.2		↓
26	NM_002121	MHC class II, DP beta 1 (DPB1)	6p21.3	L, Im	↑
27	NM_013995	Lysosomal-associated membrane protein 2 (LAMP2), transcript variant LAMP2B	Xq24		↑
28	NM_018315	F-box and WO-40 domain protein 7	4q31.3		↓
29	NM_012087	General transcription factor IIIC, polypeptide 5	9q34	TF	↓
30	NM_014257	C-type lectin domain family 4, member M (*CLEC4M*)	19p13	L, Im	↓
31	NM_006595	Apoptosis inhibitor 5	11p12–q12		↑
32	NM_017650	Protein phosphatase 1, subunit 9A	7q21.3		↑
33	NM_000791	Dihydrofolate reductase	5q11.2–q13.2		↑
34	NM_000798	Dopamine receptor D5	4p16.1		↓
35	NM_004857	A kinase anchor protein 5	14q21–24		↓
36	NM_002719	Protein phosphatase 2, regulatory subunit B, gamma isoform	14q22		↑
37	NM_018349	Multiple domains with 2 transmembrane regions 2	15g26.2		↑
38	NM_020231	x010	3q13.33		↑
39	NM_006477	RAS-related on chromosome 22, transcript variant 1	22q12.2		↓
40	NM_004057	S100 calcium binding protein G	Xp22.2		↓
41	NM_019096	GTP binding protein 2	6p21–p12		↓
42	NM_005118	Tumor necrosis factor superfamily, member 15 (TNFSF15)	9q32q	Im	↓
43	NM_006599	Nuclear factor of activated T cells 5, tonicity response, transcript variant 3[b]	16q22.1	TF, Im	↑
44	NM_003070	SW1/SNF-related matrix-associated actin-dependent regulator of chromatin, subfamily a, member 2, transcript variant 1	9p22.3	TF	↑
45	NM_002930	Ras-like without CAAX2	18q12.3		↓
46	NM_007023	Rap guanine nucleotide exchange factor 4	2q31–q32		↓
47	NM_000781	Cytochrome P-450, family 11, subfamily A, polypeptide 1	15q23–q24		↓
48	NM_001060	Thromboxane A2 receptor, transcript variant 2	19p13.3		↓
49	NM_002357	MAX dimerization protein 1	2p13–p2		↓
50	NM_003635	*N*-deacetylase/*N*-sulfotransferase 2	10q22		↓
51	NM_003220	Transcription factor AP2 alpha	6p24	TF	↓
52	NM_015239	ATP/GTP binding protein 1	9q21.33		↑

Table 1 (continued)

No.	Accession number	Protein function	Map location	Notes[b]	CD[c]
53	NM_016626	Ring finger and KH domain-containing 2	18q21.1	TF	↑
54	NM_021995	Urotensin 2, transcript variant 1	1p36		↑
55	NM_014167	HSPC128	12q21.31	L	↑
56	NM_018676	Thromboxane type I domain-containing 1, transcript variant 1	13q14.3		↑
57	NM_005387	Nucleoporin 98 kDa, transcript variant 3	11p15.5		↓
58	NM_015894	Stathmin-like 3	20q13.3		↓
59	NM_012343	Nicotinamide nucleotide transhydrogenase	5p13.1		↓
60	NM_020119	Zinc finger CCH type, antiviral 1, transcript variant 1	7q34	TF, Im	↑
61	NM_005879	TRAF interacting protein	3p21.31	Im	↓
62	NM_012175	F-box protein 3, transcript variant 1	11p13	TF	↑
63	NM_001298	Cyclic nucleotide gated channel alpha 3	2q11.2		↑
64	NM_001463	Frizzled-related protein	2qter		↑
65	NM_002544	Oligodendrocyte myelin glycoprotein	17q11.2		↑
66	NM_012177	F-box protein 5	6q25–q26		↑
67	NM_005392	PHD finger protein 2, transcript variant 1	9q22.31	TF	↓
68	NM_005760	CCAAT/enhancer binding protein zeta	2p22.2	TF	↑
69	NM_003864	Sin-3 associated polypeptide	11q34.1		↑
70	NM_001139	Arachidonate 12-lipoxygenase, 12R type	17p13.1		↓
71	NM_005809	Periredoxin 2, transcript variant 1	19p13.2		↓
72	NM_012200	Beta-1,3-glucoronyl transferase 3	11q12.3		↓
73	NM_013351	T-box 21	17q21.32	TF	↓
74	NM_000254	*S*-methyltetrahydrofolate-homocysteine methyltransferase	1q43		↑
75	NM_020402	Cholinergic receptor, nicotinic, α-polypeptide 10	11p15.5		↓
76	NM_002509	NK2 transcription factor-related, locus 2	20pter-q11.23	TF	↑
77	NM_002347	Lymphocyte antigen 6 complex, locus H	8q24.3	Im	↓
78	NM_003175	Chemokine (C motif) ligand 2	1q23–25	Im	↓
79	NM_005718	Actin-related protein 2/3 complex, subunit 4	3p25.3		↓
80	NM_006296	Vaccinia-related kinase 2	2p16–p15		↑
81	NM_002749	Mitogen-activated protein kinase 7, transcript variant 3	17p11.2		↓
82	NM_014433	Rhabdoid tumor deletion region gene 1	22q11.2		↓
83	NM_002505	Nuclear transcription factor Y, alpha, transcript variant 1	6p21.3	TF	↓
84	NM_015313	Rho guanine nucleotide exchange factor 12	11q23.3		↑
85	NM_001206	Kruppel-like factor 9	9q13	TF	↓
86	NM_024036	Leucine-rich repeat and fibronectin III-containing 4	11q13.2		↓
87	NM_018933	Protocadherin beta 13	5q31	L	↑
88	NM_024582	FAT tumor suppressor homology cadherin	4q28.1		↑
89	NM_013251	Tachykinin 3	12q13–q21	L	↑
90	NM_005712	HERV–HLTR associating 1	8q24		↓
91	NM_017623	Cyclin M3 transcript variant 1	2p12–p11.2		↑
92	NM_000171	Glycine receptor, alpha 1	5q23		↓
93	NM_003505	Frizzled homolog 1	7q21		↑
94	NM_003630	Peroxisomal biogenesis factor 3	6q23–q24		↑
95	NM_016436	PHD finger protein 20	20q11.22–q11.23	TF	↑
96	NM_007191	WNT inhibitory factor 1	12q14.3	L	↑
97	NM_001856	Collagen type XVI, alpha 1	9q21.31		↓
98	NM_007005	Transducin-like enhancer of split 4	10q11.2		↑
99	NM_006965	Zinc finger protein 11b	8q13.2	TF	↓
100	NM_021833	Uncoupling protein 1	4q28–q31		↑
101	NM_002751	Mitogen-activated protein kinase 11, transcript variant 1	22q3.33		↓
102	NM_003409	Zinc finger protein 161	18pter-p11.2	TF	↑
103	NM_024303	Zinc finger and SCAN domain-containing 5	19q13.43	TF	↓
104	NM_018048	Mago-nashi homolog	12p13.2	L	↑
105	NM_014358	C-type lectin domain family 4. member E (*CLEC4E*)	12p13.31	Im	↑
106	NM_004198	Cholinergic receptor, nicotinic, α-polypeptide 6	8p11.2		↓

Table 1 (continued)

No.	Accession number	Protein function	Map location	Notes[b]	CD[c]
107	NM_014443	Interleukin-17B (IL-17B)	5q32–34	Im	↓
108	NM_022354	Spermatogenesis-associated 1	1p22.3		↓
109	NM_030824	Zinc finger protein 442	19p13.2	L, TF	↑
110	NM_007252	POU domain, class 6, transcription factor 2	7p14–p13	TF	↓
111	NM_020389	Transient receptor, potential cation channel, subfamily 6, member 7	5q31.1	L	↓

[a]Transcripts are listed in order of decreasing magnitude of the *F*-statistic for disease state. Transcripts included all that belong to the group with FDR less than 1%
[b]*L* gene region linked to IBD, *TF* transcription factor, *Im* immune system function
[c]Significant increase (↑) or decrease (↓) in CD relative to control

ing a significant effect of disease state at the 1% level and FDR of less than 1%, 508 showed significant HSD at the 1% level between means for CD and control and FRD of less than 1%. None showed an HSD between UC and control that was significant at the 1% level. Table 1 lists all transcripts with annotated protein function and map location from the RefSeq database (Pruitt et al. 2005) that showed a significant effect of disease state at the 1% level and a significant HSD at the 1% level, with FDR of 1% or less in each case. These included 22 known or putative transcription factors and 12 genes mapping to genomic regions that have shown evidence of association with IBD (Table 1).

Discussion

An approach based on analysis of variance was applied to microarray data from a publicly available database in order to identify transcripts with significant differential expression across ileum and colon in inflammatory bowel disease (IBD). Statistically significant differences in expression levels between Crohn's disease (CD) and control were observed for numerous transcripts. Such differences were more rarely seen in the case of ulcerative colitis (UC), as is expected, since the latter is not expected to affect the ileum. By combining data from two disease states and two tissues, this approach achieved the statistical power to detect transcripts with consistently altered expression across ileum and colon in CD. Furthermore, the analysis of variance design used tissue as a block effect, thereby removing this effect statistically and increasing the power to test for effects of disease states (control, CD, and UC). This approach made it possible to extract information on gene expression changes in CD from a data set lacking independent replicates from CD-affected patients.

The magnitude of the detectable difference among disease states, as measured by the *F*-statistic for the effect of disease state, was not strongly correlated with the ratio of raw expression scores between CD and control or between UC and control. Rather, the transcripts with the highest *F*-statistics often had low ratios of disease scores to control scores, and vice versa. This surprising result evidently occurred because many of the transcripts with high ratios of disease to control were transcripts lacking a consistent pattern of expression change in disease state across ileum and colon.

While some of the latter possibly represented genes with a pattern of tissue-specific differential expression in one or both diseases, the available data did not make it possible to test statistically for a tissue-specific expression difference in most cases. On the other hand, in the case of 168 transcripts for which replicated data were available, there were no significant results in tests for tissue-by-disease-state interaction. Yet these 168 transcripts showed significant effects of disease state at rates comparable to the other transcripts, suggesting that they were not atypical of the data set as a whole. The absence of detectable tissue-by-disease-state interactions suggests that inconsistent patterns of expression between the two tissues may often have been simply due to stochastic fluctuations without biological importance.

The analysis of variance identified numerous transcripts with differential expression in CD. These included transcripts from a number of genes with known roles in regulating gene expression in signal transduction and in immune recognition, all processes likely to be involved in CD. The transcripts with significant results at the 1% level and annotated function from the RefSeq database included a number of potential interest for both the mechanism of causation and the pathology of CD (Table 1). These included 22 known or putative transcription factors, among them five zinc finger proteins (Table 1). Eleven of the 22 transcription factors showed significantly higher expression levels in CD than in the control, while 11 showed significantly lower expression levels in CD than in the control (Table 1). Two α-polypeptides from cholinergic receptors showed significantly lower expression in CD than in controls (nos. 75 and 106, Table 1).

Among the most interesting genes in Table 1 were 12 genes mapping to genomic regions that have shown evidence of linkage to IBD (Table 1). Of the seven regions with the strongest association to CD (16q12, 12p13.2–q24.1, 19p13, 1p36, 5q31, 14q11–12, and the *HLA* region

on chromosome 6), all but 16q12 and 14q11–12 are represented by one or more transcripts in Table 1. DPβ1, which was showed significantly increased expression in CD (no. 26, Table 1), maps to the *HLA* region. 19p13 included three genes with annotated function and significant evidence of differential expression in CD: the zinc finger protein 442 (no. 109, Table 1) and EDG5 (no. 9, Table 1), with increased expression in CD, and *CLEC4M* (no. 30, Table 1), with decreased expression in CD. EDG5 is a G-protein-coupled receptor involved in cell proliferation (An et al. 2000). *CLEC4M* forms part of an evolutionarily conserved cluster of type II membrane-associated C-type lectins, belonging to the CD209 family and expressed on dendritic cells (Geijtenbeck et al. 2000; Bashirova et al. 2003).

Table 1 included six genes mapping to the broad region of chromosome 12 (12p13.2–q24.1) that shows association with IBD. Among these was *BTG1* (no. 17, Table 1), which has an anti-proliferative function (Iwai et al. 2004) and showed increased expression in CD. Likewise, showing increased expression in CD was HSPC128 (no. 55, Table 1), a transcript identified from hematopoietic stem/progenitor cells (Zhang et al. 2000). TAC3 (no. 89, Table 1) encodes a protein known as tachykinin 3 or neurokinin-B that encode molecules modulating physiological processes via G-protein-coupled receptors (Pal et al. 2004), and this gene also showed evidence of increased expression in CD. Also of interest with regard to chromosome 12 linkage was *CLEC4E*, (no. 105, Table 1), which maps to 12p13.31, just outside the region associated with CD. *CLEC4E* is a close relative of *CLEC4M* on chromosome 19, but these two C-type lectin genes showed contrasting patterns in CD (Table 1). Whereas *CLEC4M* showed significantly decreased expression in CD, *CLEC4E* showed significantly increased expression in CD (Table 1).

Another functionally interesting gene mapping to a region linked with CD (1p36) is *TNFRSF4* (no. 18, Table 1), which showed significantly increased expression in CD (Table 1). The protein product, also known as CD134 and OX40, is important for T cell proliferation and is upregulated in multiple sclerosis (Kashiwakura et al. 2004; Carboni et al. 2003). Procadherin beta 13 (no. 87, Table 1), with significantly increased expression in CD, forms part of a cluster in the 5q31 region encoding members of the protocadherin family, which are involved in cell adhesion (Wu et al. 2001).

Current models of CD implicate sensing of peptidoglycan and/or other bacterial cell wall components as a key event in the causation of disease, a line of investigation encouraged by the discovery that CARD15/NOD2 is associated with CD (Girardin et al. 2003). Since CD is believed to be a complex genetic disease with at least seven susceptibility loci (Girardin et al. 2003), genes with differential expression in CD that map to regions previously showing association with CD would seem plausible candidates for further association studies. This would seem to be especially true in the case of genes that play a role in bacterial sensing or the transduction of signals from such sensing. The present analysis revealed a number of genes having these characteristics, and investigation of polymorphism at these loci may yield further insights into the mechanism of causation of CD.

Acknowledgements This research was supported by grant GM43940 from the National Institutes of Health.

References

An S, Zheng Y, Bleu T (2000) Sphingosine 1-phosphate-induced cell proliferation, survival, and related signaling events mediated by G protein-coupled receptors Edg3 and Edg5. J Biol Chem 275:288–296

Barrett T, Suzek TO, Troup DB, Wilhite SE, Ngau W-C, Ledoux P, Rudnev D, Lash AE, Fijibuchi W, Edgar R (2005) NCBI GEO: mining millions of expression profiles—database and tools. Nucleic Acids Res 33:D562–D566

Bashirova AA, Wu L, Cheng J, Martin TD, Martin MP, Benveniste RE, Lifson JD, KewalRamani VN, Hughes A, Carrington M (2003) Novel member of the *CD209* (*DC-SIGN*) gene family in primates. J Virol 77:217–227

Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Stat Soc B 57:289–300

Bouma G, Strober W (2003) The immunological and genetic basis of inflammatory bowel disease. Nat Rev Immunol 3:521–533

Carboni S, Aboul-Enein F, Waltzinger C, Killenn N, Lassmann H, Peña-Rossi C (2003) CD134 plays a crucial role in the pathogenesis of EAE and is upregulated in the CNS of patients with multiple sclerosis. J Neuroimmunol 145:1–11

Cho JH, Nicolae DL, Ramos R, Fields CT, Rabenau K, Corradino S, Brant SR, Espinosa R, LeBeau M, Hanauer SB, Bodzin J, Bonen DK (2000) Linkage and linkage disequilibrium in chromosome band 1p36 in American Chaldeans with inflammatory bowel disease. Hum Mol Genet 9:1425–1432

Dieckgraefe BK, Stenson WF, Korzenik JR, Swanson PE, Harrington CA (2000) Analysis of mucosal gene expression in inflammatory bowel disease by parallel nucleotide arrays. Physiol Genomics 4:1–11

Devauchelle V, Chiocchia G (2004) Quelle place pour les puces à AND dans les maladies inflammatoires? Rev Med Interne 25:732–739

Gasche C, Alizadeh BZ, Peña AS (2003) Genotype–phenotype correlations: how many disorders constitute inflammatory bowel disease? Eur J Gastroenterol Hepatol 15:599–608

Geijtenbeck TB, Terensma R, van Vliet SJ, van Duijnhoven GC, Adema GJ, van Kooyk Y, Figdor CG (2000) identification of DC-SIGN, a novel dendritic cell-specific ICAM-3 receptor that supports primary immune responses. Cell 100:575–585

Girardin SE, Hugot J-P, Sansonetti PJ (2003) Lessons from Nod2 studies: towards a link between Crohn's disease and bacterial sensing. Trends Immunol 24:652–658

Heller RA, Schena M, Chai A, Shalon D, Bedilion T, Gilmore J, Woolley DE, Davis RW (1997) Discovery and analysis of inflammatory disease-related genes using cDNA microarrays. Proc Natl Acad Sci U S A 94:2150–2155

Hugot JP, Chamaillard M, Zouali H, Lesage S, Cezard JP, Belaiche J et al (2001) Association of NOD2 leucine-rich repeat variants with susceptibility to Crohn's disease. Nature 411:599–603

Iwai K, Hirata K, Ishida T, Takeuchi S, Hirase T, Rikitake Y, Kojima Y, Inoue N, Kawashima S, Yokoyama M (2004) An anti-proliferative gene *BTG1* regulates angiogenesis in vitro. Biochem Biophys Res Comm 316:628–635

Kashiwakura J, Yokoi H, Saito H, Okayama Y (2004) T cell proliferation by direct crosstalk between OX40 ligand on human mast cells and OX40 on human T cells: comparison of gene expression profiles beetween human tonsillar and lung cultured mast cells. J Immunol 173:5247–5257

Kok K, Stokkers P, Reitsma PH (2004) Genomics and proteomics: implications for inflammatory bowel diseases. Inflamm Bowel Dis 10(Suppl 1):S1–S6

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

Mannick EE, Bonomolo JC, Horswell R, Lentz JJ, Serano M-S, Zapata-Velandia A, Gastanaduy M, Himel JL, Rose SL, Udall JN Jr, Hornick CA, Liu Z (2004) Gene expression in mononuclear cells from patients with inflammatory bowel disease. Clin Immunol 112:247–257

Negoro K, McGovern DPB, Kinouchi Y, Takahashi S, Lench NJ, Shimosegawa T, Carey A, Cardon LR, Jewell DP, van Heel DA (2005) Analysis of the IBD5 locus and potential gene–gene interactions in Crohn's disease. Gut 52:541–546

Pal S, Nemeth MJ, Bodine D, Miller JL, Svaren J, Thein SL, Lowry PJ, Bresnick EH (2004) Neurokinin-B transcription in erythroid cells. J Biol Chem 279:31348–31356

Pruitt KD, Tausova T, Maglott DR (2005) NCBI reference sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. Nucleic Acids Res 33:D501–D504

Russell RK, Nimmo ER, Satsangi J (2004) Molecular genetics of Crohn's disease. Curr Opin Genet Dev 14:264–270

Sokal RR, Rohlf FJ (1981) Biometry, 2nd edn. Freeman, San Francisco, CA

Van Heel DA, Fisher SA, Kirby A, Daly MJ, Rioux JD, Lewis CM, Genome Scan Meta-Analysis Group of the IBD International Genetics Consortium (2005) Inflammatory bowel disease susceptibility loci defined by genome scan meta-analysis of 1952 affected relative pairs. Hum Mol Genet 13:763–770

Watts DA, Satsangi J (2002) The genetic jigsaw of inflammatory bowel disease. Gut 50:31–36

Wu Q, Zhang T, Cheng J-F, Kim Y, Grimwood J, Schmutz J, Dickson M, Noonan JP, Zhang MQ, Myers RM, Maniatis T (2001) Comparative DNA sequence analysis of mouse and human protocadherin gene clusters. Genome Res 11:389–404

Zhang Q-H, Ye M, Wu X-Y, Ren S-X, Zhao M, Zhao C-J, Fu G, Shen Y, Fan H-Y, Lu G, Zhong M, Xu X-R, Han Z-G, Zhang J-W, Tao J, Huang Q-H, Zhao J, Hu G-X, Gu J, Chen S-J, Chen Z (2000) Cloning and functional analysis of cDNAs with open reading frames for 300 previously undefined genes expressed in CD34+ hematopoietic stem/progenitor cells. Genome Res 10: 1546–1560