Pediatric Crohn disease patients exhibit specific ileal transcriptome and microbiome signature

Yael Haberman,[1] Timothy L. Tickle,[2,3] Phillip J. Dexheimer,[4] Mi-Ok Kim,[5] Dora Tang,[1] Rebekah Karns,[4] Robert N. Baldassano,[6] Joshua D. Noe,[7] Joel Rosh,[8] James Markowitz,[9] Melvin B. Heyman,[10] Anne M. Griffiths,[11] Wallace V. Crandall,[12] David R. Mack,[13] Susan S. Baker,[14] Curtis Huttenhower,[2,3] David J. Keljo,[15] Jeffrey S. Hyams,[16] Subra Kugathasan,[17] Thomas D. Walters,[11] Bruce Aronow,[4] Ramnik J. Xavier,[3,18,19] Dirk Gevers,[3] and Lee A. Denson[1]

[1]Division of Gastroenterology, Hepatology and Nutrition, Cincinnati Children's Hospital Medical Center, Cincinnati, Ohio, USA. [2]Department of Biostatistics, Harvard School of Public Health, Boston, Massachusetts, USA. [3]Broad Institute of Massachusetts Institute of Technology and Harvard University, Cambridge, Massachusetts, USA. [4]Biomedical Informatics and [5]Division of Biostatistics and Epidemiology, Cincinnati Children's Hospital Medical Center, Cincinnati, Ohio, USA. [6]The Children's Hospital of Philadelphia, Department of Pediatrics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, Pennsylvania, USA. [7]Department of Pediatrics, Division of Gastroenterology, Medical College of Wisconsin, Milwaukee, Wisconsin, USA. [8]Goryeb Children's Hospital/Atlantic Health, Morristown, New Jersey, USA. [9]Division of Pediatric Gastroenterology and Nutrition, Cohen Children's Medical Center of New York, New Hyde Park, New York, USA. [10]Department of Pediatrics, UCSF, San Francisco, California, USA. [11]Division of Gastroenterology, Hepatology and Nutrition, Hospital for Sick Children, University of Toronto, Toronto, Ontario, Canada. [12]Division of Pediatric Gastroenterology, Hepatology and Nutrition, Nationwide Children's Hospital, Columbus, Ohio, USA. [13]Department of Pediatrics, Children's Hospital of Eastern Ontario, Ottawa, Ontario, Canada. [14]Digestive Diseases and Nutrition Center, Department of Pediatrics, University at Buffalo, Buffalo, New York, USA. [15]Division of Pediatric Gastroenterology, Children's Hospital of Pittsburgh, Pittsburgh, Pennsylvania, USA. [16]Division of Digestive Disease and Nutrition, Connecticut Children's Medical Center, Hartford, Connecticut, USA. [17]Division of Gastroenterology, Hepatology and Nutrition, Department of Pediatrics, Emory University, Atlanta, Georgia, USA. [18]Gastrointestinal Unit and Center for the Study of Inflammatory Bowel Disease and [19]Center for Computational and Integrative Biology, Massachusetts General Hospital, Harvard Medical School, Boston, Massachusetts, USA.

Interactions between the host and gut microbial community likely contribute to Crohn disease (CD) pathogenesis; however, direct evidence for these interactions at the onset of disease is lacking. Here, we characterized the global pattern of ileal gene expression and the ileal microbial community in 359 treatment-naive pediatric patients with CD, patients with ulcerative colitis (UC), and control individuals. We identified core gene expression profiles and microbial communities in the affected CD ilea that are preserved in the unaffected ilea of patients with colon-only CD but not present in those with UC or control individuals; therefore, this signature is specific to CD and independent of clinical inflammation. An abnormal increase of antimicrobial dual oxidase (*DUOX2*) expression was detected in association with an expansion of Proteobacteria in both UC and CD, while expression of lipoprotein *APOA1* gene was downregulated and associated with CD-specific alterations in Firmicutes. The increased *DUOX2* and decreased *APOA1* gene expression signature favored oxidative stress and Th1 polarization and was maximally altered in patients with more severe mucosal injury. A regression model that included *APOA1* gene expression and microbial abundance more accurately predicted month 6 steroid-free remission than a model using clinical factors alone. These CD-specific host and microbe profiles identify the ileum as the primary inductive site for all forms of CD and may direct prognostic and therapeutic approaches.

Introduction

Current evidence suggests that the inflammatory bowel diseases (IBDs) Crohn disease (CD) and ulcerative colitis (UC) are caused by a complex interaction among host genetic background, microbial shifts, and environmental cues, leading to inappropriate chronic activation of the mucosal immune system (1–3). While it is difficult to establish causality in patient-based studies, it is reasonable to suggest that large inception cohorts that include clinical, genetic, mucosal, and microbial profiling might be the optimal way to address the diversity associated with IBD pathogenesis with adequate power. A recent meta-analysis of IBD genetic studies identified 163 IBD risk loci (4), and many of these risk alleles exhibit infection-related balancing natural selection (4). Consistent with this finding, an overall dysfunction in the human gut microbial community has been described in both longstanding adult-onset IBD (5) and treatment-naive pediatric-onset IBD (6), and patients exhibit altered responses to bacterial DNA (7). Animal models have conclusively shown causality in the requirement for bacterial colonization in the development of intestinal inflammation in genetically susceptible hosts (3). However, characterization of host/microbial profiles in the affected and unaffected mucosa at the onset of disease in large patient-based cohorts has been lacking.

Diagnostic and therapeutic decisions in IBD are based primarily on clinical and endoscopic severity and histopathologic analysis of intestinal biopsies. With this approach, only a minority of patients experience durable remission, which may be due to substantial heterogeneity in the underlying pathogenic mechanisms not accounted for by current classification systems (8). The two main forms of IBD, CD and UC, share many genetic

Conflict of interest: The authors have declared that no conflict of interest exists.
Submitted: January 29, 2014; **Accepted:** May 29, 2014.
Reference information: *J Clin Invest*. 2014;124(8):3617–3633. doi:10.1172/JCI75436.

Table 1. RISK RNA-seq cohort clinical and demographic characteristics

	Ctl (*n* = 43)	UC1 (*n* = 45)	cCD1 (*n* = 37)	UC2 (*n* = 28)	cCD2 (*n* = 26)	iCD1 (*n* = 90)	iCD2 (*n* = 90)	All iCD (*n* = 180)	iCD-DU (*n* = 78)	iCD-noDU (*n* = 102)
Mean (SD) age (yr)	11 (3)	12 (3)	12 (4)	13 (4)	13 (3)	12 (3)	12 (3)	12 (3)	12 (3)	12 (3)
Male gender (%)	65	47	54	71	54	63	60	62	58	65
MED ethnicity (3 of 4 grandparents) (%)	97	87	91	86	77	90	88	89	91	88
Perianal involvement (%)	0	0	16	0	24	19	17	18	19	17
Ileal deep ulcers (%)	0	0	0	0	0	44	42	43	100	0
BMI z-score < −2 (%)	3	2	22	4	15	28	16	22	23	21
PCDAI at diagnosis										
≤10 (inactive, %)	NA	NA	8	NA	4	8	10	9	12	7
11–30 (mild, %)	NA	NA	36	NA	38	37	44	41	32	48[A]
>30 (moderate to severe, %)	NA	NA	56	NA	58	55	46	50	56	46

Differences between selected groups were tested by ANOVA for continuous variables and χ^2 for dichotomous variables. MED, mixed European descent.
[A]$P = 0.0421$ vs. iCD-DU.

susceptibility loci but differ in anatomical location and disease behavior, often dictating different medical and surgical approaches, further implying the need for additional classification approaches. More importantly, disease classification should take into account potential response to current and future therapy. This is especially relevant now, when second-line biologic medications are on track for approval (9, 10) by the FDA, and there is an ongoing need to identify patients that will derive the greatest relative benefit from early biological therapy with anti–TNF-α agents (8, 11). A recent comparative effectiveness study that used the same inception cohort as the current report showed that early anti–TNF-α therapy is superior to other approaches in achieving 1-year steroid- and surgery-free remission (SSFR) (11). No clinical or demographic parameters were specifically associated with therapeutic response, suggesting that additional information is needed to better define patient subsets. The use of gene expression or microbial markers to support diagnosis and adjust therapy for specific subsets of IBD is currently limited. In one single-center study, researchers tested for association between ileal or colonic gene expression and subsequent clinical and mucosal response to anti–TNF-α therapy in adults with established CD (12). Overall, 12 of 19 patients experienced healing of colonic ulcers with anti–TNF-α, and baseline expression of specific colonic genes was associated with subsequent response. Remarkably, only 1 of 18 patients experienced healing of ileal ulcers with anti–TNF-α, and hence no gene pattern was defined for this response. Therefore, there is a pressing clinical need to define pathogenic mechanisms driving ileal ulcers and to test whether genomic and microbial data will improve patient classification.

Patient-based and murine studies have suggested that the terminal ileum, which contains approximately 50% of the Peyer's patches in the gut, plays a central role as a sensor of bacterial colonization and tolerance and is likely the primary inductive site for mucosal immunologic pathogenesis in CD (13–15). We therefore focused on capturing the net output of contributing environmental and genetic factors by performing high-throughput ileal microbial community and host gene expression analyses from DNA and RNA extracted from ileal biopsies collected at the time of patients' initial diagnostic endoscopy (Figure 1A). Previous studies have been limited by small sample size, variable biopsy locations, long duration of disease, and multiple prior treatments (12, 16–19). We addressed these limitations by analyzing ileal biopsies obtained at the time of diagnosis in a large treatment-naive prospective inception cohort of patients with early-onset (pediatric) IBD, known as the RISK cohort. We have recently used the RISK cohort to define (6) a mucosal microbial axis in treatment-naive CD, which includes an increased abundance of Enterobacteriaceae, Pasteurellaceae, Veillonellaceae, and Fusobacteriaceae and decreased abundance of Erysipelotrichales, Bacteroidales, and Clostridiales. This microbial axis correlated strongly with clinical disease severity within CD. Additionally, we found that antibiotics, sporadically prescribed in response to the initial symptoms, amplified the microbial dysbiosis (6). Here, we further expand our understanding, at the ileal mucosal level, by specifically characterizing the host gene expression and the associated microbial abundance in a broader spectrum of disease phenotypes. To capture a CD-specific signature, independent of the confounding effects of local mucosal inflammation, we defined the global pattern of gene expression and microbial community in both clinically affected (ileal CD [iCD]) and unaffected (colon-only CD [cCD]) ilea, in comparison with noninvolved ilea in UC and non-IBD controls (Ctl). We aimed to thereby identify both core CD pathogenic host/microbe associations and to test for association among these host-transcriptomic and microbial factors, disease severity, and subsequent clinical course (Figure 1A).

Results

Identification of a core iCD gene expression signature. The Crohn's and Colitis Foundation of America–sponsored RISK study is a prospective inception cohort study, which enrolled 1,276 pediatric patients with IBD at diagnosis at 28 sites in North America between 2008 and 2012 (Supplemental Figure 1 and Supplemental Excel file 1; supplemental material available online with this article; doi:10.1172/JCI75436DS1). All patients were treatment naive, with ileal biopsies obtained during the initial diagnostic colonoscopy. Only subjects with a confirmed diagnosis of CD, UC, or non-IBD

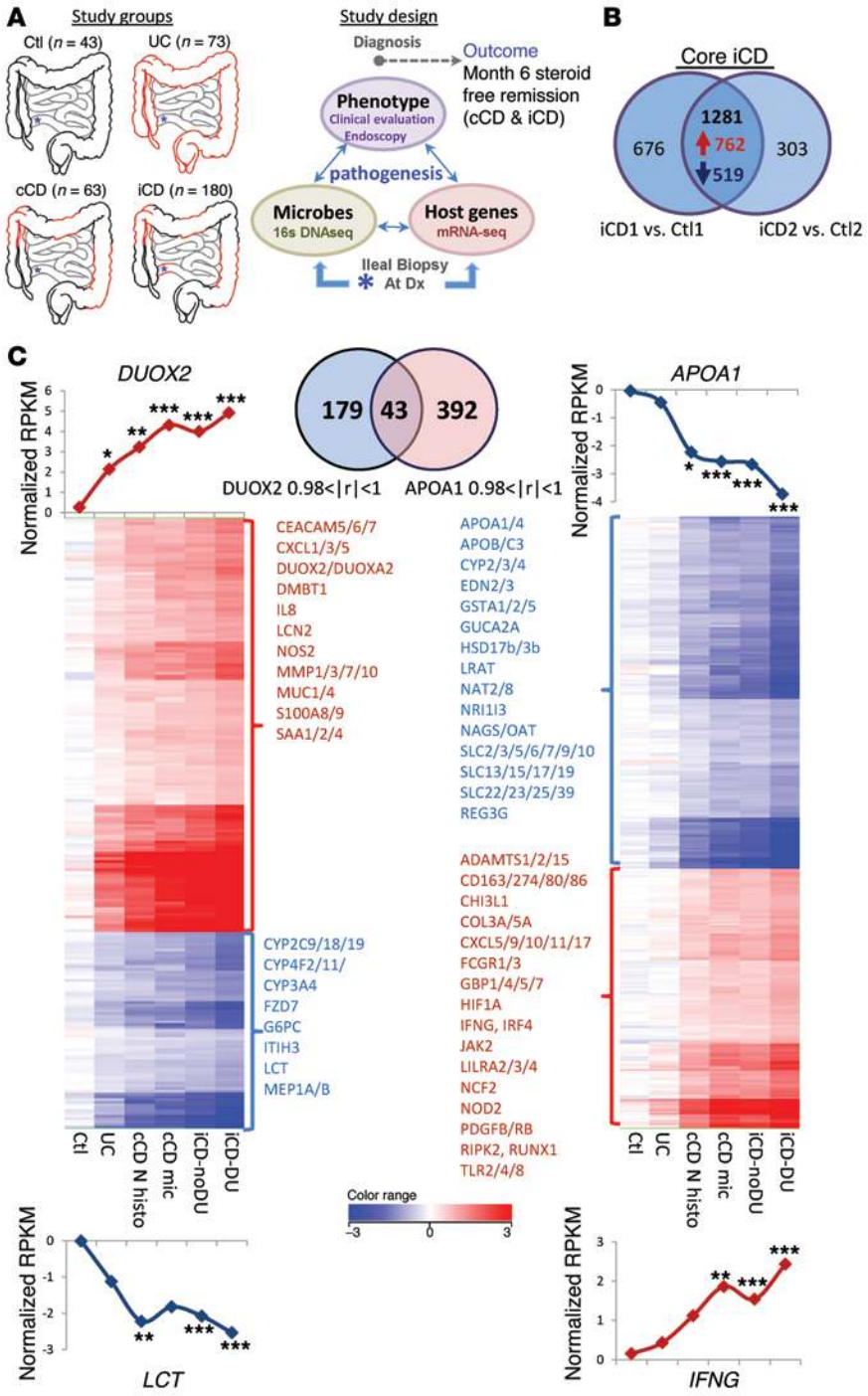

Figure 1. *APOA1* and *DUOX2* gene coexpression signatures define pathogenic ileal modules. (**A**) Study groups included Ctl, UC, cCD, and iCD. Ileal 16S DNA sequencing and mRNA sequencing were used to define microbial profiles and gene expression signatures, respectively. (**B**) Venn diagram shows 1,281 genes (core iCD signature) that were differentially expressed (fold change ≥ 1.5) between 2 independent iCD and Ctl groups. (**C**) Heat maps of average gene expression for the clinical subgroups for *DUOX2* or *APOA1* gene coexpression signatures (Pearson correlation $0.98<|r|<1$). Venn diagram indicates *DUOX2* and *APOA1* gene coexpression signature overlap. Selected upregulated (red) and downregulated (blue) genes are listed, with the graphs showing the average *DUOX2*, *LCT*, *APOA1*, and *IFNG* gene expression across groups. Differences between patient subgroups were tested using Kruskal-Wallis with Dunn's multiple comparison test. *$P < 0.05$, **$P < 0.01$, ***$P < 0.001$ vs. Ctl. cCD-mic, cCD with abnormal histological features; cCD N histo, cCD with normal histology.

Ctl during follow-up were included (Table 1 and Supplemental Excel file 2). Our initial analyses identified a core iCD gene expression signature composed of 1,281 genes that were differentially expressed in the ilea of 2 independent iCD groups compared with Ctl (Figure 1B, Supplemental Figure 2A, and Supplemental Excel file 3). We conducted functional annotation enrichment analyses using several data sources, including gene ontology to map groups of related genes within the core iCD gene signature to upstream regulators, immune cell types, pathways, phenotypes, and biologic functions (Supplemental Figure 2, B and C, and Supplemental Excel files 4–6). These analyses provided a way to identify which known transcriptional regulators, biologic processes, and immune cell types were likely to be upregulated or downregulated within the ileum based on a given gene expression pattern. Analyses were conducted using ToppGene (20), ToppCluster (21), and Ingenuity Pathway Analysis software. The relative degree of upregulation or downregulation of a given transcriptional regulator or biologic process was provided by the activation *z* score obtained as an output from Ingenuity Pathway Analysis software (Supplemental Figure 2B), while *P* values for the specific cell type, pathway, phenotype, and biologic function associated with upregulated or downregulated genes were obtained as an output from ToppGene (Supplemental Excel files 4–6). Ileal enrichment for a given immune cell class was illustrated as shown in Supplemental Figure 2C by colored bars on the *x* axis, with the significance for each individual cell type within the class shown as the $-\log_{10}$ (*P* value) on the *y* axis.

The core iCD signature was enriched for genes induced by bacterial products and proinflammatory cytokine signaling, including the Th1 cytokine IFN-γ, whereas genes induced by several nuclear receptors, including HNF4α, were suppressed (Supplemental Figure 2B). Functional analyses (20) identified enrichment of innate antimicrobial responses and an unexpected profound loss of nuclear recep-

Table 2. Top upregulated differentially expressed genes within the core iCD gene signature

	Gene	iCD1 FC	iCD2 FC
1	*DUOXA2*	34.3	51.9
2	*MMP3*	21.4	29.7
3	*AQP9*	21.2	32.8
4	*IL8*	18.6	27.8
5	*DUOX2*	15.8	19.2

Genes are listed 1–5, with that with the most differential expression being first. FC, fold change.

tor-dependent lipid metabolic functions (Supplemental Figure 2B and Supplemental Excel files 4 and 5). Immune cell enrichment analysis was most significant for genes expressed by granulocytes, myeloid dendritic cells, macrophage/monocytes, and lymphoid stromal cells (Supplemental Figure 2C and Supplemental Excel file 6). Within the top 5 most upregulated genes, we noted the epithelial antimicrobial dual oxidase *DUOX2* and its maturation factor *DUOXA2*. Within the downregulated genes, we noted the antiinflammatory HNF4α-dependent lipoprotein *APOA1* as the top downregulated gene (Tables 2 and 3).

The core iCD signature contains APOA1 and DUOX2 gene coexpression signatures. Strong experimental linkage to gut inflammation (22–25), together with the high fold-change expression differences, prompted us to specifically examine the pattern of *APOA1* and *DUOX2* gene expression in the following clinically defined subgroups: Ctl, UC, cCD, and iCD. Patients in the cCD subgroup met diagnostic criteria for CD but lacked visible ileal inflammation on endoscopy; patients with UC also had disease limited to the colon. The cCD subgroup was further divided into those with normal ileal histopathology and those with microscopic ileal inflammation. The iCD subgroup was further divided into 2 groups based upon the presence of deep ulcers, iCD with deep ulcers [iCD-DU], and iCD without deep ulcers [iCD-noDU]) (26). Figure 1C shows progressive increased expression of *DUOX2* in the ileum across the spectrum of Ctl, UC, cCD, and iCD, while suppression of *APOA1* expression was specific to all forms of CD. Importantly, the 2 cCD groups showed a similar pattern of gene expression as the 2 iCD groups; this pattern was different from UC and Ctl (Figure 1C). We next performed Pearson correlation analysis to define other genes within the core 1,281 iCD signature with similar patterns of gene expression as *DUOX2* or *APOA1* across the predefined patient groups that therefore had high likelihood for coregulation and shared biologic function. Pearson coexpression correlation analysis ($0.98<|r|<1$) of the core iCD genes with either *APOA1* or *DUOX2* ileal expression identified 435 genes and 222 genes, respectively (Figure 1C, Supplemental Figures 3 and 4, and Supplemental Excel files 7 and 8). The heat map with specific indicated upregulated and downregulated genes further shows the expression pattern within the 2 gene coexpression signatures across the indicated patient groups (Figure 1C). Genes within the *DUOX2* gene coexpression signature showed increasing or decreasing signal intensity across the spectrum of patient groups from Ctl to iCD-DU (Figure 1C), while genes within the *APOA1* gene coexpression signature showed alterations that were specific to all forms of CD. Interestingly, the mature enterocyte digestive enzyme lactase (*LCT*) and members of the mucin family (e.g., *MUC4*) were included in the *DUOX2* gene coexpression signature (Figure 1C and Supplemental Figures 3 and 4). By comparison, induction of the Th1 cytokine *IFNG* and its downstream effector chemokine (*CXCL9*) was included within the CD-specific *APOA1* gene coexpression signature (Figure 1C and Supplemental Figures 3 and 4). Importantly, genes within the 2 gene coexpression signatures showed similar expression patterns between cCD with normal or abnormal histology and iCD-noDU and therefore were independent of overt clinical inflammation (Figure 1C).

Immunohistochemistry confirmed the expected apical localization of DUOX2 within villous enterocytes in the CD ileum, in conjunction with suppression of epithelial APOA1 and evidence of increased lipid peroxidation (Figure 2A). We next performed functional annotation enrichment analyses of the *DUOX2* and *APOA1* gene coexpression signatures using Ingenuity Pathway Analysis, ToppCluster (21), and Cytoscape (ref. 27 and Figure 2B). Functional annotation enrichment analyses for the *DUOX2* gene coexpression signature essentially captured the key features of the NF-κB–dependent upregulated innate antimicrobial response contained within the core iCD signature and reduction of metabolic processes regulated by the glucocorticoid nuclear receptor NR3C1 (Supplemental Excel file 9 and Supplemental Figure 3C). Functional annotation enrichment analyses of the *APOA1* gene coexpression signature revealed decreased expression of genes regulated by HNF4α, PPARγ, and several other nuclear receptors, with associated predicted decreases in several lipid metabolic and antioxidant functions (Supplemental Excel file 9 and Supplemental Figure 4C). Importantly, these changes occurred in association with upregulation of IFN-γ/STAT1-dependent genes involved in Th1 polarization. Collectively, these results defined 2 core *DUOX2* and *APOA1* gene coexpression signatures within the CD ileum, regulating both enterocyte and innate and adaptive immune functions.

The APOA1 coexpression genes are enriched within a CD-specific signature that is independent of clinical inflammation and may be used for patient classification. Remarkably, the majority of the 1,281 genes that comprised the core iCD gene signature were also differentially expressed in the ilea of patients with cCD compared with Ctls (1,055 of 1,281 genes, 82%) as opposed to only 18% (232 of 1,281) that were differentially expressed in the ilea of patients with UC compared with Ctls (Supplemental Figure 5A). This included upregulation of *DUOX2* and downregulation of *APOA1* in the

Table 3. Top downregulated differentially expressed genes within the core iCD gene signature

	Gene	iCD1 FC	iCD2 FC
1	*APOA1*	−10.8	−6.6
2	*NAT8*	−10.6	−6.8
3	*AGXT2*	−10.2	−6.7
4	*CUBN*	−9.4	−6.9
5	*FAM151A*	−9.4	−8.3

Genes are listed 1–5, with that with the least differential expression being first.

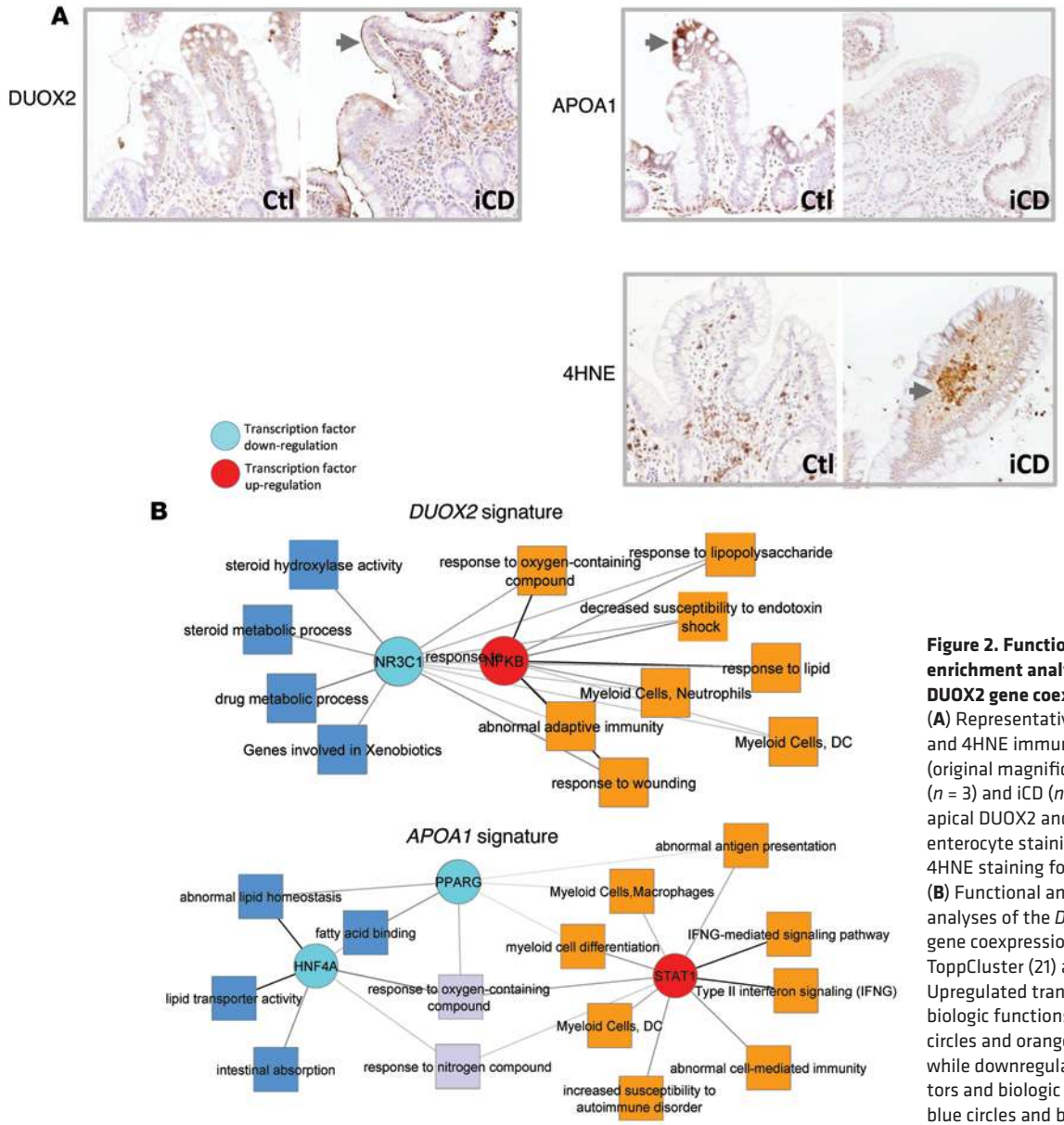

Figure 2. Functional annotation enrichment analysis of the APOA1 and DUOX2 gene coexpression signatures. (**A**) Representative ileal DUOX2, APOA1, and 4HNE immunohistochemistry (original magnification, ×40) for Ctl (n = 3) and iCD (n = 7). Arrows indicate apical DUOX2 and intracellular APOA1 enterocyte staining and lamina propria 4HNE staining for lipid peroxidation. (**B**) Functional annotation enrichment analyses of the *DUOX2* and *APOA1* gene coexpression signatures using ToppCluster (21) and Cytoscape (27). Upregulated transcription factors and biologic functions are shown as red circles and orange boxes, respectively, while downregulated transcription factors and biologic functions are shown as blue circles and boxes, respectively.

cCD ileum. We then asked whether the *APOA1* gene coexpression signature would be enriched within the cCD ileum and whether this genomic information could be used for patient classification. Indeed, a large portion (572 of 614; 93%) of the genes from the combined *APOA1* and *DUOX2* gene coexpression signatures were also differentially expressed between cCD and Ctl (Supplemental Excel file 10 and Supplemental Figure 5B). In contrast, only 15% (89 of 614) of the genes from the combined *APOA1* and *DUOX2* gene coexpression signatures, primarily from the *DUOX2* signature, were differentially expressed between UC and Ctl (Supplemental Excel file 11 and Supplemental Figure 5B). Unsupervised hierarchical clustering analysis identified groups of biopsies with similar ileal gene expression profiles; this analysis tested whether the cCD ileal transcriptional profile for these 2 gene coexpression signatures would cluster with the iCD transcriptional profile, whereas the UC transcriptional profile would cluster with the non-IBD Ctl transcriptional profile. Results of this analysis showed that, for the *APOA1* gene coexpression signature, most cCD ileal biopsies clustered together with iCD ileal biopsies, while most UC ileal biopsies clustered with non-IBD Ctls (Supplemental Figure 4, $\chi^2 = 7.7$, $P = 0.005$). A similar clustering was not observed when using the *DUOX2* gene coexpression signature (Supplemental Figure 3, $\chi^2 = 1.1$, $P = 0.3$). Collectively, these data demonstrate that altered transcriptional profiles in CD were observed even in the histologically normal cCD ileum.

Arriving at the correct diagnosis in patients with IBD with colon-only involvement (i.e., differentiating cCD from UC) can be challenging with current clinical tools. Hypothesizing that diagnosis could be enhanced by genomic information, in addition to our training cCD and UC groups (cCD1 and UC1), we used an independent validation (cCD2 and UC2) cohort and tested classification of these groups using both unsupervised and supervised approaches. To further refine a CD-specific signature in the ileum, we identified differentially expressed genes (179 with fold

Table 4. Characteristics of regression models to predict 6-month SSFR

	Clinical variables only	Clinical variables and gene expression	Clinical, gene expression, and microbial variables
AIC	215.293	212.706	208.089
BIC	233.929	234.447	239.148
C statistics (or AUC)	0.705	0.721	0.760
Likelihood ratio test	Clinical variables vs. clinical variables and gene expression	0.0269	
	Clinical variables vs. clinical, gene expression, and microbial variables	0.0043	

AIC, Akaike's information criterion; BIC, Bayesian information criterion.

change ≥ 2 and 93 with fold change ≥ 2.5) between the cCD and UC training cohorts (cCD1 and UC1), which were also contained in the *DUOX2* or *APOA1* gene coexpression signatures, whereby 82% (147 of 179) were from the *APOA1* gene coexpression signature (Figure 3A and Supplemental Excel file 12). Unsupervised hierarchical clustering analysis to test for similarities in ileal gene expression between groups showed that both the training and validation cCD cohorts clustered with iCD, irrespective of microscopic inflammation, in contrast to the UC training and validation cohorts, which clustered with Ctl (Figure 3B and Supplemental Figure 5C, $\chi^2 = 11.8$, $P = 0.0006$ for training cohort).

Supervised classification algorithms may be used to develop models to classify unknown patient groups (in this case to classify cCD vs. UC) based upon gene expression data that differ within a training set of known patient groups. The Support Vector Machine algorithm is a machine learning approach, which is commonly used to classify 2 patient classes based upon differential expression of biologic data and, in fact, was used recently to distinguish cCD from UC with 77% accuracy by cross-validation in a model based upon differential colon protein abundance (28). We therefore conducted a supervised classification analysis using the cCD1 gene list versus UC1 gene list for the cCD1/UC1 training cohort to develop a classification model using the Support Vector Machine algorithm in Avadis and then tested the accuracy of the model on the independent validation cohort (26 cCD2 and 28 UC2). Applying the model to the independent validation cohort resulted in accurate classification of 41 of 54 patients (76% overall accuracy). Among 17 patients classified by the model as cCD, 15 carried a clinical diagnosis of cCD (88% accuracy). However, among 37 patients classified by the model as UC, only 26 carried a clinical diagnosis of UC (70% accuracy). Of the patients with cCD misclassified as UC, this included 4 with microscopic ileal histologic involvement, 6 with normal ileal histology, and 1 whose ileal histology was not recorded. This result demonstrated a reasonable level of accuracy in using the ileal gene expression data alone to classify the 2 colon-only forms of IBD.

We next performed functional annotation enrichment analyses for the 179 genes specifically altered in CD independent of ileal inflammation (cCD vs. UC) (Figure 3C). Genes from the *DUOX2* gene coexpression signature (Supplemental Excel file 12) showed decreased CYP3A4-related NR3C1 glucocorticoid receptor signaling and an increase in NF-κB–dependent innate immune activation and tissue degradation genes. Of particular interest, within the *APOA1* gene coexpression signature (Supplemental Excel file 12), a marked downregulation of several HNF4α and PPARγ nuclear receptor-dependent glutathione s-transferases (*GSTA1*) and apolipoproteins (*APOA1/4*, *APOB*, and *APOC3*) with antioxidant function and an upregulation of an IFN-γ/STAT1-dependent Th1 signature was again noted. Collectively these results define a core iCD gene coexpression signature and associated biologic functions that were largely independent of overt clinical inflammation (Figure 3C).

Upstream regulators and biologic functions associated with mucosal ulceration. Endoscopic severity in CD is defined by the presence of deep ulcers (Figure 4A), and FDA approval of new therapies requires evidence of mucosal healing (8, 29). However, the host expression signature and microbial composition associated with the pathogenesis of this critical clinical parameter has not been defined in a large cohort of treatment-naive patients. We identified 345 differentially expressed genes associated with deep ulcers within the core iCD signature (Figure 4A, Supplemental Figure 6, and Supplemental Excel file 13). Analysis performed using Ingenuity Pathway Analysis software identified enrichment of genes associated with increased bacterial products and proinflam-

Table 5. Multiple regression analysis, including clinical, gene expression, and microbial variables

		P value	OR	CI
Age ≥ 10 yr vs. < 10 yr		0.8868	0.944	0.430, 2.075
iCD-DU vs. iCD-noDU	PCDAI > 30	0.6244	0.771	0.271, 2.188
	PCDAI ≤ 30	0.0029	4.713	1.701, 13.057
Anti-TNF therapy		0.0020	5.181	1.828, 14.706
APOA1 expression level > 80th percentile		0.0152	3.058	1.241, 7.576
Blautia abundant (>70th percentile) vs. nonabundant	*Veillonella* abundant	0.5183	1.634	0.368, 7.25
	Veillonella nonabundant	0.0028	0.231	0.089,0.604

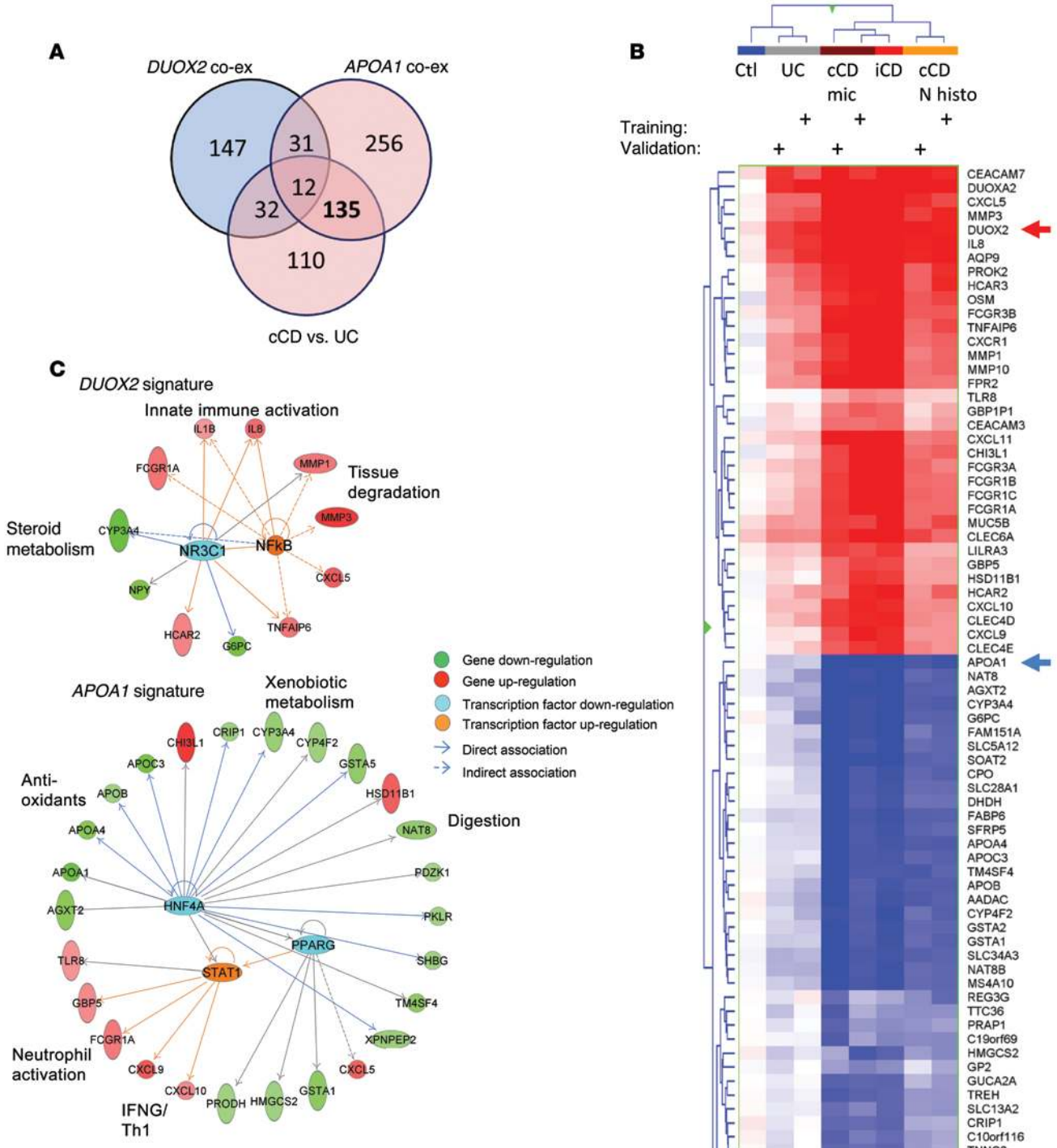

Figure 3. The ileal APOA1 gene coexpression signature is specific to all forms of CD. (**A**) The Venn diagram shows the overlap of 179 genes differentially expressed 2-fold between patients with cCD and UC and genes within the *DUOX2* and *APOA1* coexpression modules. (**B**) A heat map of 93 of the above-mentioned 179 genes after averaging and hierarchical clustering of gene expression, with a fold change of 2.5 between the cCD and UC training cohorts for the indicated clinical subgroups. The training cohort included 17 cCD with abnormal histological features, 17 cCD with normal histology, and 45 UC. The independent validation cohort included 14 cCD with abnormal histological features, 11 cCD with normal histology, and 28 UC. The blue arrow indicates *APOA1*, and the red arrow indicates *DUOX2*. (**C**) The results of functional annotation enrichment analyses (IPA, Ingenuity Systems) using the *DUOX2* and *APOA1* gene coexpression signatures. Upregulated transcription factors and target genes are shown in orange and red, respectively, while downregulated transcription factors and target genes are shown in blue and green, respectively. Biologic functions associated with these groups of genes are also shown.

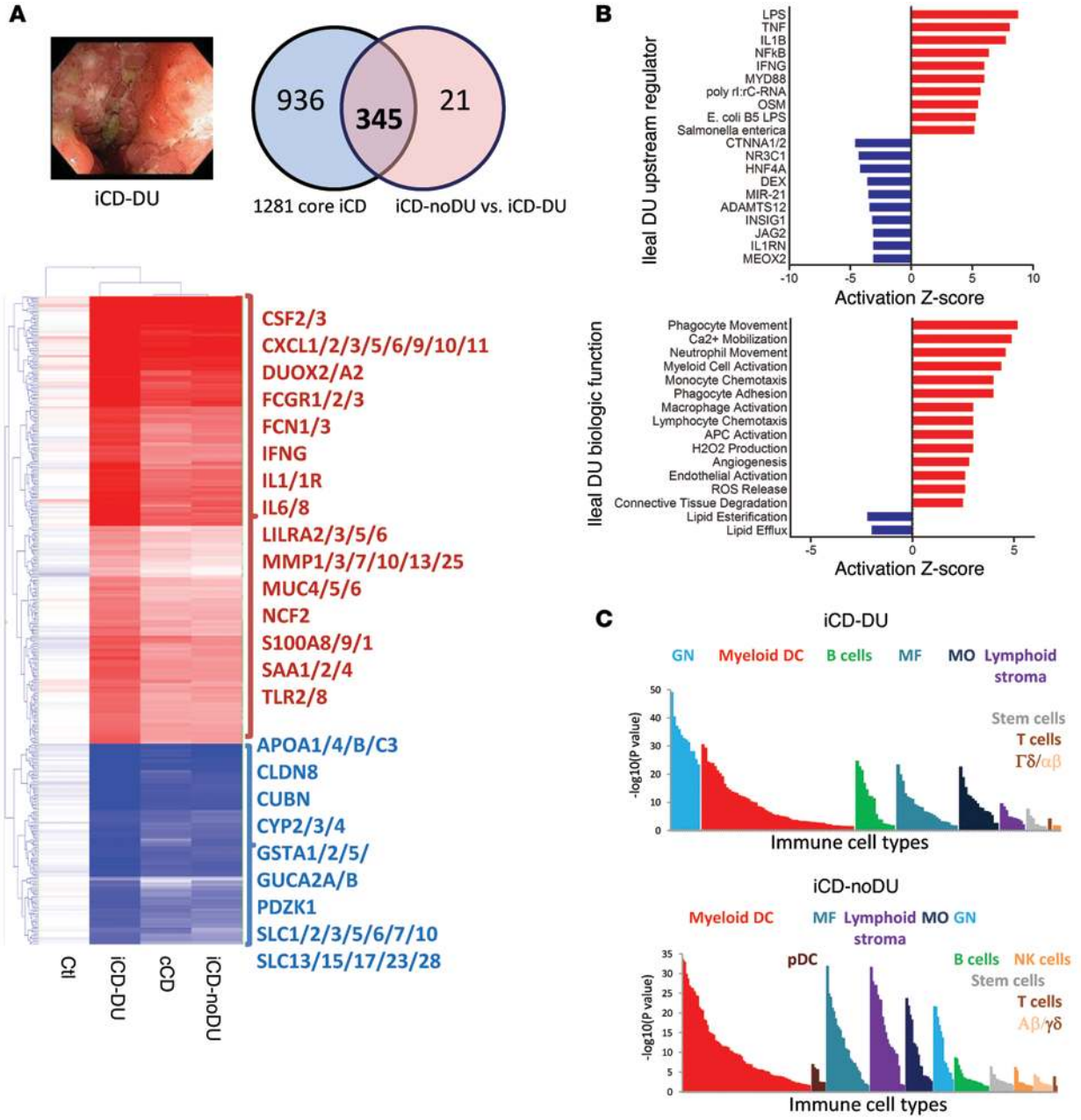

Figure 4. Gene expression signature and biologic pathways associated with mucosal ulceration. (**A**) Image of ileal deep ulcers and Venn diagram showing 345 genes from the core iCD gene signature, which overlap with differentially expressed genes (Audic Claverie method with Benjamini-Hochberg FDR correction [0.05], fold change ≥ 1.5), between iCD-DU and iCD-noDU. A heat map of averaged gene expression and that after hierarchical clustering for the iCD-DU gene list for the indicated clinical subgroups. Specific upregulated (red) and downregulated (blue) genes are listed. (**B**) IPA functional annotation enrichment analyses to detect upstream regulators and biologic functions associated with the mucosal ulceration (iCD-DU) gene expression signature are shown. The activation *z* score for biologic function enrichment (*P* value range: 6E-8 to 3E-26) and upstream regulator enrichment (*P* value range: 3E-10 to 1E-56) depicts the degree of activation or suppression of a given regulator or function. (**C**) Immune cell–type enrichment of upregulated genes for iCD-DU (238 of 345 genes) and upregulated genes for iCD-noDU (524 of 936 genes) was determined using the Immunological Genome Project data series as a reference through ToppGene (20). Ileal enrichment for a given immune cell class (e.g., granulocytes [GN]) is illustrated by colored bars on the *x* axis, with the significance for each individual cell subtype within the class shown as the $-\log_{10}$ (*P* value) on the *y* axis. pDC, plasmacytoid DC; MF, macrophages; MO, monocytes.

matory cytokine signaling, reduced nuclear receptor signaling, and a broad range of immune responses, including myeloid cell and lymphocyte activation (Figure 4B). Several biologic functions, including reactive oxygen species and hydrogen peroxide production, angiogenesis, and connective tissue degradation, were notably enriched within the iCD-DU gene coexpression signature (Figure 4B). We tested for immune cell–type enrichment between iCD-DU and iCD-noDU by conducting functional annotation enrichment analyses of immune cell–type gene expression using the Immunological Genome Project data series as a reference through

ToppGene (20). Ileal enrichment for a given immune cell class (e.g., granulocytes) is illustrated in Figure 4C by colored bars on the *x* axis, with the significance for each individual cell subtype within the class shown as the $-\log_{10}(P \text{ value})$ on the *y* axis. The most significant signal within iCD-DU was for granulocytes, followed by myeloid dendritic cells and B cells, while iCD-noDU showed a different order of enrichment, with high representations of myeloid dendritic cells, macrophages, and lymphoid stromal cells, in the absence of the pronounced granulocyte signal (Figure 4C). These data define biologic pathways and associated immune cell types specifically associated with severe mucosal ulceration.

A fundamental clinical problem, as well as a difficulty in trial design in CD, is the lack of correlation between clinical disease activity indices and the severity of mucosal injury as defined by deep ulcers (8, 29). Consistent with this observation, stratifying the 180 patients with iCD by the pediatric Crohn's disease activity index (PCDAI) identified only 43 differentially expressed genes within the core iCD signature between those with mild symptoms (PCDAI ≤ 30) and those with moderate-severe symptoms (PCDAI > 30) at diagnosis. Therefore, there was no clear association between the modest number of ileal genes associated with more severe symptoms and the much more robust gene coexpression signature associated with tissue ulceration (Supplemental Excel files 13 and 14).

Shifts in the ileal microbial community are preserved in the unaffected CD ileum. Murine studies have established a definitive requirement for bacterial colonization in the development of mucosal inflammation in the genetically susceptible host (30, 31). Functional annotation enrichment analyses of the core iCD gene signature demonstrated enrichment of genes associated with myeloid cell responses (Supplemental Excel file 6, $P < 1.60E\text{-}44$ to $6.74E\text{-}69$), response to lipopolysaccharides (Supplemental Excel file 4, $P < 1.5E\text{-}30$), altered susceptibility to infection (Supplemental Excel file 4, $P < 3.1E\text{-}24$), and response to biotic stimulus (Supplemental Excel file 4, $P < 5.03E\text{-}38$). Remarkably, a significant proportion of genes whose ileal expression is known to be regulated by bacterial colonization of mice were contained within the *APOA1* and *DUOX2* gene coexpression signatures (Supplemental Figure 7 and ref. 15). This group included upregulated key drivers of the innate (*DUOX2* and *DUOXA2*) and adaptive (*CXCL9* and *CXCL10*) immune responses both in iCD and following bacterial colonization of the mouse ileum. This finding suggested that the host gene coexpression signal in CD may be due, in part, to shifts in the ileal microbial community. We therefore next defined the ileal microbial community within the patient cohort and tested for association with genes from the *DUOX2* and *APOA1* gene coexpression signatures.

We first performed a univariate analysis (LDA Effect Size) to test for association between ileal microbial composition and disease state in 240 subjects with CD, 163 non-IBD Ctl, and 56 subjects with UC from the RISK cohort, with no recent exposure to antibiotics (Supplemental Excel file 15). The output of this univariate analysis is illustrated by a cladogram, which demonstrates the relationship between the different bacterial taxa and disease state (Figure 5A). These results are consistent with prior studies of patients with IBD with established disease (32) and the recently published ileal, rectal, and fecal microbial community characterization of the RISK cohort (6). We observed prominent increased abundance of the Firmicutes phyla within the non-IBD Ctls that was preserved in the UC ileum but not in CD ileum, along with a shift toward Fusobacteria, Gemellaceae, and Proteobacteria expansion in the CD ileum.

Microbial shifts that have previously been defined in the ileal mucosa of patients with adult-onset CD with longstanding disease could be secondary to adaption to the diseased micro-environment or a primary event independent of local inflammation (5, 32). To address this fundamental pathogenic question, we next uniquely characterized the ileal microbial community in patients with CD without clinical ileal involvement (cCD), compared with those with overt ileal inflammation (iCD), for taxa previously shown to be associated with treatment-naive pediatric CD (6). As shown in Figure 5B, a very similar microbial shift was observed in the ilea between patients with iCD and cCD, irrespective of histologic involvement in cCD. This included expansion of Veillonellaceae, Pasteurellaceae, Neisseriaceae, Gemellaceae, and Fusobacteriaceae and persistent suppression of Lachnospiraceae, Bifidobacteriaceae, Clostridiales, and Erysipelotrichaceae in all forms of CD, while expansion of Enterobacteriaceae was not apparent in patients with no histological or clinical inflammation. To examine the possibility that local nearby cecal inflammation contributed to the microbial shift identified in the ileum, we stratified patients with cCD based on the presence of macroscopic cecal inflammation reported during colonoscopy (Supplemental Figure 8). Over all, there was persistent reduction in Lachnospiraceae, Bifidobacteriaceae, Clostridiales, and Erysipelotrichaceae in all forms of CD, with expansion of Veillonellaceae, Pasteurellaceae, Neisseriaceae, Gemellaceae, Fusobacteriaceae, and Enterobacteriaceae independent of cecal involvement. Collectively, these results defined a core CD microbial shift that occurred, like the *APOA1* and *DUOX2* gene coexpression signatures, largely in the absence of overt clinical inflammation (Figure 5B).

Ileal Firmicutes and Proteobacteria taxa abundance are associated with the APOA1 and DUOX2 gene coexpression signatures and clinical outcomes. To test for an association between the core CD host gene coexpression signatures and these specific microbial shifts, we performed multivariate analysis by linear models (MaAsLin) (5, 6) in a subgroup of patients for whom both RNA-Seq and microbial community profiling had been performed (195 CD, 50 UC, and 34 non-IBD Ctl). We specifically tested for associations between members of the ileal microbial community and representative genes from the *APOA1* (*APOA1*, *CXCL9*) and *DUOX2* (*DUOXA2*, *MUC4*, *LCT*) gene coexpression signatures, clinical group (Ctl, UC, CD), endoscopic severity (ileal deep ulcers), and clinical severity (PCDAI), while controlling for age, gender, BMI, and risk allele carriage in *NOD2*, *FUT2*, and *ATG16L1*. We were able to identify 70 significant associations between gene expression and microbial taxa and 34 significant associations between clinical parameters and microbial taxa ($P < 0.05$ and $q < 0.25$, Supplemental Figure 9 and Supplemental Excel file 16).

We then used a novel biplot approach (Figure 6A, ordination was rotated by CD group) to visualize the covariation among the ileal microbial community structure, clinical group, CD clinical disease severity (PCDAI), and ileal gene expression from MaAsLin. Sample and microbial feature coordinates were generated as a

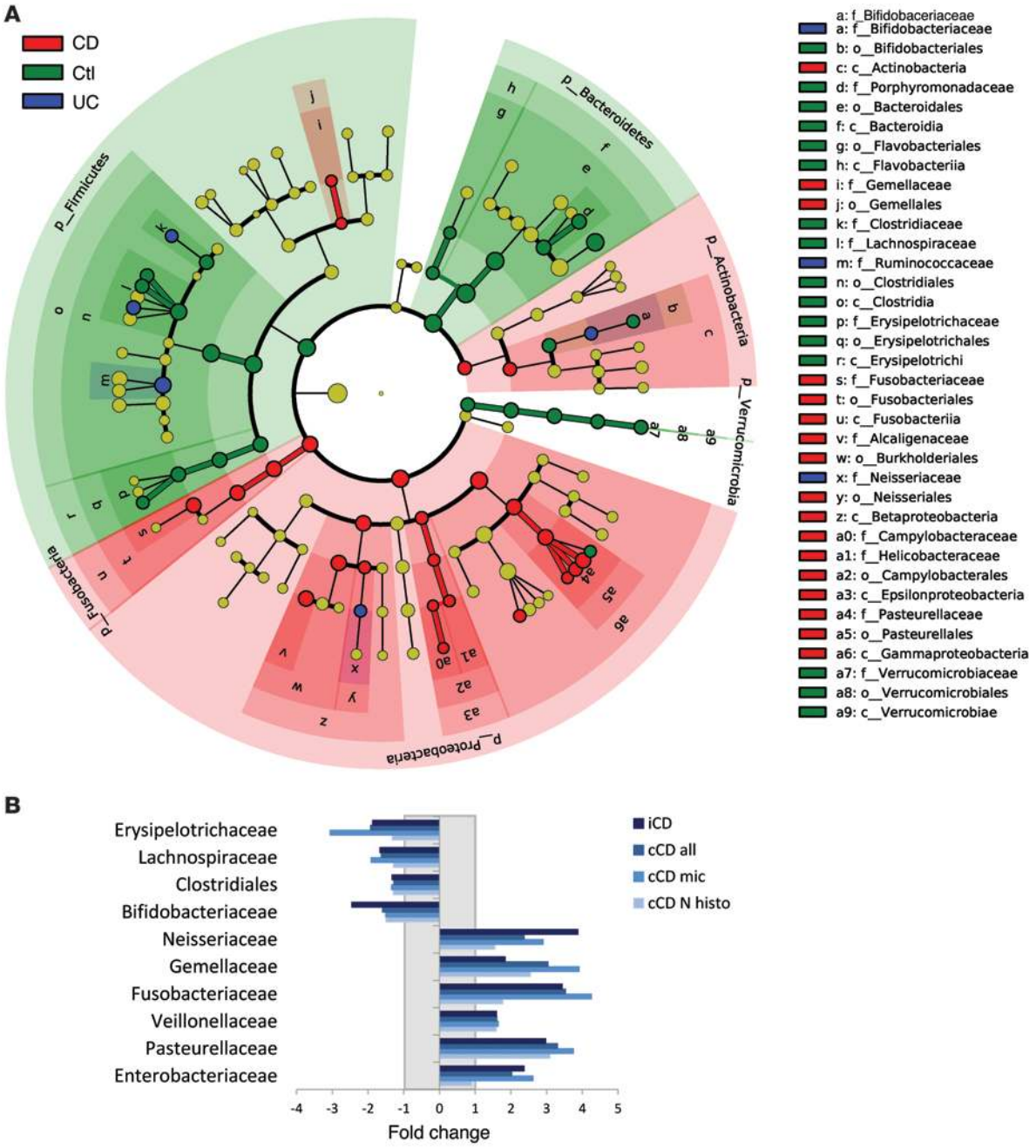

Figure 5. The ileal microbial community in patients with IBD and Ctls. (**A**) Univariate analysis (LDA Effect Size) was performed to test for association between ileal microbial composition and disease state in 240 CD (48 cCD, 178 iCD, and 14 CD with no indicated ileal involvement), 163 non-IBD Ctl, and 56 UC subjects from the RISK cohort that had not received antibiotics prior to endoscopy. The cladogram illustrates the output of this univariate analysis by demonstrating the relationship between the different bacterial taxa and disease state. Colored nodes from the center to the periphery represent marked phylum (p), class (c), order (o), family (f), genus (g), and species (s) differences detected between groups for Ctls (green), CD (red), and UC (blue). Only phylum, class, order, and family levels are indicated on the right side of the cladogram. (**B**) Fold change for each taxon was calculated by dividing the mean abundance in the cases (cCD [48 patients] or iCD [178 patients]) by that of the Ctls (154 patients) and is shown for microbiota with differential abundance in CD compared with Ctl by univariate analysis. The cCD group was further subdivided to cCD with abnormal histological features (25 patients) and cCD with normal histology (18 patients).

standard biplot, with an additional dimension of clinical and gene expression metadata. The biplot used nonmetric multidimensional scaling (NMDS) to depict the multidimensional relationship between these parameters in 2 dimensions, with the stress measurement indicating the goodness of fit of the 2-dimensional representation of the data. A stress measurement of ≤0.2 is regarded as a good fit; during the development of the biplot, 19 of 20 stress measurements were ≤0.2, demonstrating that the model was not overfitted to the data. Points were used to represent samples, labels to represent selected significant microbiota, and labeled arrows to represent clinical and molecular metadata (see Methods section for more details). In the biplot, when we used specifically

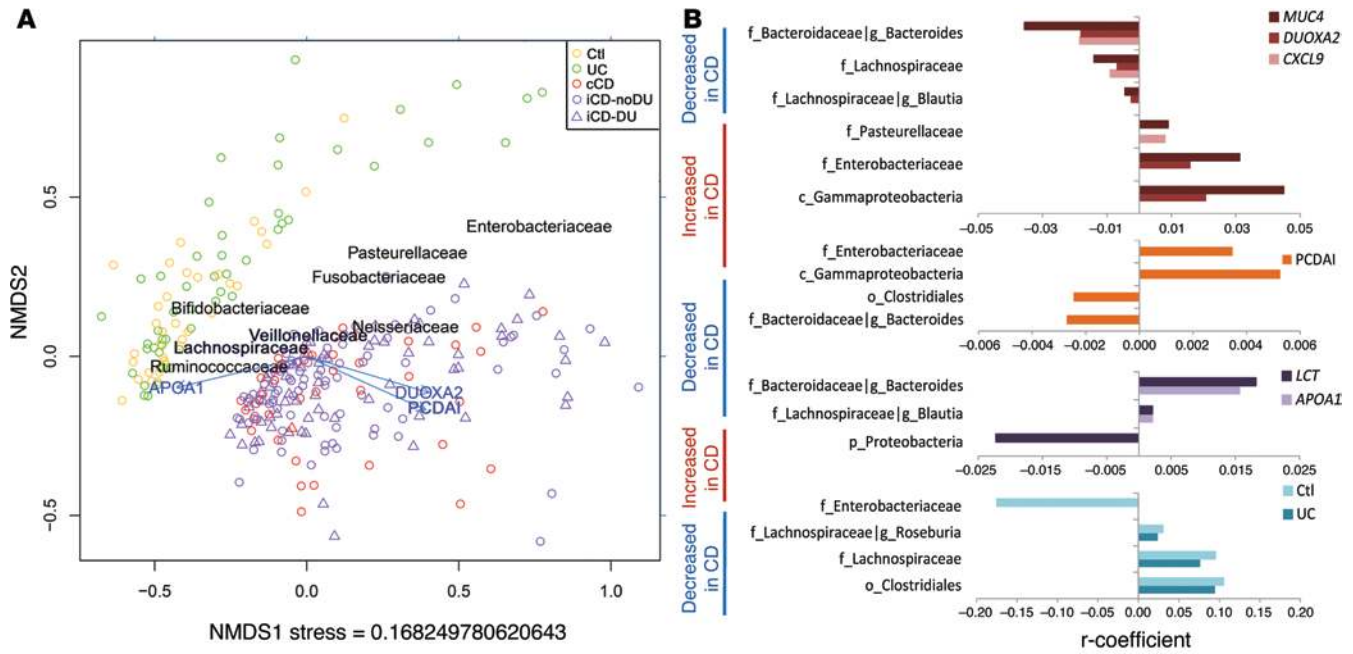

Figure 6. Covariation of the ileal microbial community structure with ileal gene expression and clinical subgroup and severity. (**A**) The biplot depicts covariation of the ileal microbial community structure with clinical group, clinical disease activity (PCDAI), and ileal gene expression using NMDS rotated by CD and based on significant associations obtained from MaAsLin. Patients with IBD and healthy CtIs are plotted as orange circles (Ctl), green circles (UC), red circles (cCD), purple circles (iCD-noDU), or purple triangles (iCD-DU). Arrows indicate the direction of covariation for *APOA1* or *DUOXA2* gene expression and clinical severity (PCDAI). Microbial taxa are labeled with capitalized lettering. The stress measurement tests for the goodness of fit of the biplot 2-dimensional depiction of the multidimensional data, with stress < 0.2 regarded as a good fit. (**B**) The bar plot graphs show effect sizes (*r* coefficients) on the *x* axes for significant associations between microbial taxa and the indicated ileal gene expression, clinical disease activity (PCDAI), and clinical groups obtained from MaAsLin, while controlling for age, gender, BMI, antibiotic exposure, and *NOD2*, *FUT2*, and *ATG16L1* risk allele carriage ($P < 0.05$, $q < 0.25$ were considered significant). Representative genes are from the *APOA1* (*APOA1* and *CXCL9*) and *DUOX2* (*DUOXA2*, *MUC4*, and *LCT*) gene coexpression signatures. The comprehensive presentation of these data with *P* values, *q* values, and taxon abundance can be found in Supplemental Excel file 16.

significant microbial taxa associations from the multivariate analysis starting at the family level, we found that Ctl and most UC ileal samples clustered together with Bifidobacteriales, Lachnospiraceae, and Ruminococcaceae and with a higher level of *APOA1* gene expression. Increasing *DUOXA2* expression and increasing clinical disease activity (PCDAI score) were represented by a vector directed to the right of the biplot in association with covariation in the ileal microbial community structure. A more central CD subgroup clustered together with Firmicutes Veillonellaceae and Betaproteobacteria Neisseriaceae, whereas the remaining CD samples and a small group of UC samples exhibited covariation with Gammaproteobacteria Pasteurellaceae, Fusobacteriaceae, and Gammaproteobacteria Enterobacteriaceae, further illustrating the dysbiosis of these microbiota in the IBD ileum. Consistent with the univariate analysis, the cCD samples were distributed across the entire spectrum of the CD microbial community structure.

We next focused on the significant specific associations that were detected between ileal gene expression and microbial abundance and between clinical metadata (mucosal ulceration, clinical disease severity, and clinical group) and microbial abundance, as shown in Figure 6B and Supplemental Excel file 16. Given our multivariate approach, the *r* coefficients shown are the effect sizes for a given microbial taxa, which represent the additional association of that microbe's relative abundance with a given gene's expression or clinical metadata, accounting for the association of all other microbes measured. Because the absolute abundance of each of the microbial taxa is relatively low, it would be anticipated that the effect sizes would also be low. *DUOXA2* and *MUC4* expression was positively associated with taxa from the Proteobacteria phylum ($q < 0.0159$ and $q < 0.0026$, respectively) and Enterobacteriaceae family, and *MUC4* and *CXCL9* expression was positively associated with taxa from the Gammaproteobacteria class Pasteurellaceae family ($q < 0.1345$ and $q < 0.176$, respectively) (Figure 6B and Supplemental Excel file 16). Additionally, CXCL9 expression showed positive association with taxa from Firmicutes Veillonellaceae ($q < 0.115$). In contrast, *DUOXA2*, *CXCL9*, and *MUC4* expression was negatively associated with taxa from the Firmicutes ($q < 0.0156$, $q < 0.137$, and $q < 0.0012$, respectively) and Bacteroidetes phyla ($q < 0.0175$, $q < 0.143$, and $q < 0.0025$, respectively) (Figure 6B and Supplemental Excel file 16).

In contrast to prior studies of patients with CD with longstanding disease, the presence of more severe mucosal injury (deep ulcers) that we have now shown to be highly associated with specific gene signatures and pathways was not associated with a significant microbial abundance shift in the multivariate analysis. However, clinical severity scoring (PCDAI) that showed no significant association with mucosal gene expression signature was specifically and significantly associated with reduction in abundance of Firmicutes ($q < 0.1$) and Bacteroidetes ($q < 0.1$) (Figure 6B) and increase of Proteobacteria taxa ($q < 0.038$). In addition, prominent preservation of several taxa within the Firmicutes phyla, including the *Roseburia* from the Lachnospiraceae family and Erysipelo-

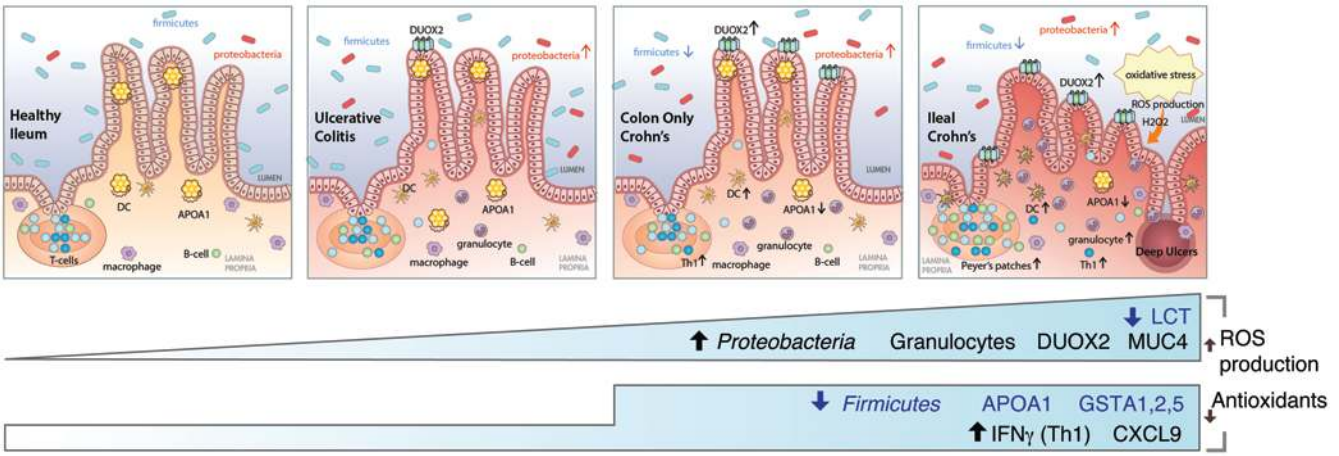

Figure 7. Host gene expression and microbial shifts across the spectrum of ileal IBD. Progressive induction of an ileal *DUOX2* host gene coexpression signature in association with expansion of Proteobacteria taxa was observed in UC, cCD, and iCD, with the greatest change relative to CtIs detected in iCD-DU (far right). By comparison, alteration of an ileal *APOA1* host gene coexpression signature in association with reduction in Firmicutes taxa was specific to all forms of CD, whereby the majority of the molecular signature of iCD was present in cCD and hence was largely independent of the degree of local inflammation. This emphasizes the central role of the ileum in the pathogenesis of all forms of CD. Maximal overall alteration of these microbial shifts and host responses favored oxidative stress and Th1 polarization and was associated with the most severe tissue injury manifested as ileal deep ulcers (far right).

trichaceae family, were observed within the UC ileum ($q < 0.0025$ and $q < 0.14$, respectively) (Figure 6B, bottom panel). By comparison, *APOA1* and *LCT* expression was positively associated with Firmicutes ($q < 0.04$ and $q < 0.12$, respectively) and Bacteroidetes ($q < 0.06$ and $q < 0.053$, respectively) taxa, while *LCT* expression was also inversely associated with abundance of specific Proteobacteria taxa ($q < 0.1$, Figure 6B, bottom panel). Collectively these data demonstrate that microbial shifts, which were detected even in the unaffected ilea of patients with cCD, were associated with specific changes in expression of genes from the *APOA1* and *DUOX2* gene coexpression signatures and the spectrum of ileal mucosal inflammation and injury in CD and UC, while reduction of taxa from the Firmicutes phyla, like the *APOA1* signature, was specific to CD and preserved in the ilea of patients with UC.

Finally, we asked whether these gene expression and microbial data, in combination with clinical data at diagnosis, would improve a prediction model for SSFR 6 months after diagnosis, when compared with a model based only on clinical factors and subsequent treatment exposures. Tables 4 and 5 present the results of a multivariable regression analysis to test the accuracy of models for 6-month SSFR, which, in an iterative fashion, included clinical parameters only, clinical and gene expression (*DUOX2* and *APOA1*) parameters, or clinical, gene expression, and microbial (taxa identified by MaAsLin) parameters. As shown in Table 4, the accuracy of each of the 3 models was assessed by the area under the curve (AUC) for a receiver operator curve analysis, with the likelihood ratio test used to formally test for a difference in accuracy between the model based only on clinical parameters and the model which included clinical, gene expression, and microbial parameters. These data demonstrated that a model that included clinical, gene expression, and microbial factors was superior to a model that included only clinical factors, with the highest AUC of 0.760 and $P = 0.0043$ (likelihood ratio test), compared with the model based on only clinical factors.

The results of the multivariable regression analysis, which included clinical, gene expression, and microbial parameters, are shown in Table 5. Neither age at diagnosis nor clinical disease activity (PCDAI) alone was associated with 6-month SSFR. Among patients with mild clinical disease activity (PCDAI ≤ 30), the presence of ileal deep ulcers was associated with a higher likelihood of achieving 6-month SSFR, as evidenced by a higher odds ratio (OR), which reached statistical significance ($P = 0.0029$). As expected, the highest OR for achieving 6-month SSFR was associated with anti-TNF therapy. After accounting for clinical and mucosal severity and anti-TNF therapy, variable selection and the classification and regression tree analysis identified higher *APOA1* expression (above the sample 80th percentile) and microbial taxa, including Lachnospiraceae *Blautia* and Veillonellaceae *Veillonella* abundance as significant prognostic factors. The relative abundance of *Blautia* and *Veillonella* interactively affected the odds of attaining 6-month SSFR: while an increased abundance of *Blautia* (above the sample 70th percentile) was negatively associated with the odds of achieving 6-month SSFR, its effect was abrogated by an increased abundance of *Veillonella* (above the sample 80th percentile). Finally, we used 10-fold cross-validation to test the overall reliability of the model and the potential for overfitting. We divided the data set into 10 subsets and used each subset as a validation set, while using the remaining 9 as a test set. Cross-validation showed that the clinical outcome prediction model using clinical, genetic, and microbial data performed reliably: the predictive power measured by the AUC on the validation sets was 0.777 on average, with SEM = 0.011. These data demonstrated that a model which includes baseline *APOA1* expression and *Blautia* and *Veillonella* abundance accurately predicts 6-month clinical outcome.

Discussion

Current evidence suggests that host/microbe interactions have played a fundamental role in the rise of multifactorial autoim-

mune diseases across the globe (33–36). Here, we report results of combined clinical, genetic, and high-throughput mucosal transcriptomic and microbial community profiling at the time of IBD diagnosis of treatment-naive patients, as part of a large multicentered, North American pediatric IBD inception cohort study. Using this comprehensive approach, we provide evidence for novel core microbial shifts and associated host gene expression profiles that are present at the onset of disease and are largely independent of overt clinical inflammation. Our results support the existence of a pathogenic model along the spectrum for ileal IBD (UC, cCD, iCD), as illustrated in Figure 7. We detected a progressive alteration of an innate antimicrobial *DUOX2* coexpression gene signature in association with expansion of Proteobacteria taxa across all forms of IBD (UC, cCD, and iCD), whereby mature enterocyte digestive functions, including *LCT* expression, decreased as *DUOX2* expression increased. We noted an additional CD-specific alteration, already present in the unaffected ilea of patients with cCD, involving suppression of an antiinflammatory/antioxidant *APOA1* gene coexpression signature, in conjunction with induction of an adaptive *IFNG/CXCL9* Th1 signature and depletion of certain Firmicutes and Bacteroidetes taxa. Collectively these bacterial shifts and host responses favor oxidative stress and Th1 polarization and are further amplified in association with more severe mucosal injury, as seen in the presence of mucosal ulceration (Figure 7, iCD-DU).

Our findings of substantially increased expression of both *DUOXA2* and *DUOX2* in IBD suggest a central role for DUOX2 in this setting. Indeed, DUOX2 interacts with the CD risk gene *NOD2* in generating intestinal epithelial cell responses to bacterial products (23), and susceptibility to spontaneous colitis in mice with deletion of antioxidant glutathione peroxidase function has been mapped to a locus containing *Duox2* (24). Altered lipoprotein composition and associated oxidative stress were previously noted in pediatric CD (25), and a prior report of the global pattern of ileal gene expression in longstanding adult-onset CD also noted profound suppression of *APOA1* and *APOAC3* gene expression (19). In fact, clustering analysis of the ileal biopsies in the prior adult-onset study showed that differentiation between CD and Ctl samples was driven by downregulated genes involved in organic acid and lipid metabolic processes and solute/cation transporter activity. This is quite similar to the results of the clustering analysis performed in the current study using the *APOA1* gene coexpression signature and provides further evidence that this is a central pathogenic pathway in CD. Remarkably, *APOA1*, *APOA4*, and *APOC3* reside in close proximity on chromosome 11 and were all downregulated within the *APOA1* gene coexpression signature, suggesting regulation by common transcriptional factors. Indeed, HNF4α and HNF4γ binding has been described within the *APOA1/A4/C3* chromosomal region (Supplemental Figure 10), and intestinal epithelial cell–specific deletion of *Hnf4a* leads to spontaneous intestinal inflammation in mice (37).This substantial suppression of the antioxidant lipoproteins in CD can hence serve as a potential target for future therapies, whereby supplementing the apolipoprotein axis could be beneficial to reverse some of the CD-specific pathogenesis in the ileum.

The broad shifts in Proteobacteria and Firmicutes we observed were remarkably similar to those previously reported in the ilea of adult patients with IBD with longstanding disease (5, 32). This finding suggests that, at least in a subset of patients, these microbial shifts may be a stable feature of refractory disease. However, our study adds significantly to the current existing knowledge, whereby we specifically show that ileal dysbiosis exists in IBD in the absence of overt clinical inflammation, potentially implying that such dysbiosis is not simply the result of the inflammatory process per se, as was previously suggested (5, 32). Consistent with this idea and with the understanding that microbial compositions affect host gene expression and vice versa (38), we observed high concordance between alteration of genes from both the *DUOX2* and *APOA1* gene coexpression signatures and ileal gene expression detected in germ-free mice following bacterial colonization (ref. 15 and Supplemental Figure 7).

We therefore applied a previously established multivariate approach (MaAsLin) (5, 6) to test for association between clinical metadata and microbial taxa to assess, for the first time, associations between selected genes from the *DUOX2* and *APOA1* signatures and microbial taxa, while controlling for host *ATG16L1*, *NOD2*, and *FUT2* genotype. Our analyses showed a significant association between expression of components of the *APOA1* module and specific Firmicutes and Bacteroidetes taxa and components of the *DUOX2* module and specific Proteobacteria taxa and Firmicutes taxa. Of particular interest was the association among the abundance of Proteobacteria Pasteurellaceae, Firmicutes Lachnospiraceae, and genes associated with both the innate (*DUOXA2* and *MUC4*) and adaptive (*IFNG/CXCL9*) immune responses.

While the overall *r* coefficients or effect sizes that they represent were relatively low for specific gene/microbe associations, the associated *P* and *q* values were highly significant. Each of the effect sizes for a given microbial taxa represented the additional association of that microbe's relative abundance with a given gene's expression, accounting for the association of all other microbes measured. Because the absolute abundance of each of the tested microbial taxa was relatively low, it would be anticipated that the effect sizes would also be low. However, we do show that significant separation on the biplot between CD and healthy (Ctl) and disease (UC) Ctls was achieved by including these more specific taxa, and these genus level taxa were also shown to improve a model for 6-month remission. Therefore, these results suggest that although the taxa at the genus and species level show a low ileal abundance and hence low effect size for association with gene expression, their presence is more specifically associated with clinical IBD subgroup and outcome and therefore pathogenesis. Mechanistic animal studies, including mono-association of germ-free mice with specific microbes, will be required to define the direct strength of association between these microbes and ileal gene expression.

A fundamental goal of the current study was to test whether knowledge of the ileal gene expression and/or microbial profile at diagnosis would have utility in patient classification beyond the current clinical system. In fact, application of a supervised classification approach to the ileal gene signature which distinguished cCD from UC in a training cohort was able to classify an independent validation cohort with reasonable accuracy. This may guide future genomic classification systems for colon-only IBD. We were further able to test for an association between the net host/microbe

association present at diagnosis and 6-month SSFR. As expected, the highest odds of achieving 6-month SSFR were associated with exposure to anti–TNF-α therapy, with only a modest effect of baseline clinical or mucosal severity. After accounting for these clinical and treatment factors, we found that patients with CD with higher baseline ileal *APOA1* gene expression and increased *Veillonella* abundance relative to *Blautia* were more likely to attain 6-month SSFR. This finding may have implications for novel therapeutic approaches targeting the *APOA1* gene coexpression signature and these specific microbes and stratification of early anti–TNF-α therapy to patients least likely to achieve 6-month remission.

Our study has several strengths but also some limitations. We examined combined host-transcriptomic and microbial profiles within ileal biopsies obtained at the time of diagnosis in a large multicenter sample of treatment-naive pediatric patients with IBD. We included both disease and non-IBD Ctls and used highly sensitive sequencing and novel analytic approaches in the largest cohort reported to date. We were thereby able to make several novel observations regarding the core iCD transcriptome and associated microbiota not previously reported, while avoiding the potential confounding effects of prior therapy, longstanding duration of disease, and bias toward analysis of surgical resection specimens in prior reports (16, 17, 32). However, while we report several associations between gene coexpression modules, specific microbiota, and clinical and mucosal severity, these studies in patients cannot conclusively establish causality. Future studies will need to test these putative mechanisms in animal models in order to define primary pathways driving mucosal gene expression and disease penetrance.

In summary, our results have several pathogenic and clinical implications. Endoscopic appearance of the classic inflammatory process was absent in the ilea of patients with UC and cCD, while iCD-associated gene expression and microbial community differences were remarkably preserved within the cCD subgroup. We defined an ileal gene set primarily from the *APOA1* gene coexpression signature that is CD specific and can potentially be used to differentiate cCD from UC. Gene-expression based classifiers are currently used in oncology (39) but are not available in the IBD field. Our suggested tissue-based molecular classifications may challenge the accepted IBD clinical classification and pave the way for potential future diagnostic tissue-based genomic and microbial classifications. In agreement with the clinical discordance between mucosal ulcers and clinical severity (PCDAI), we failed to identify a common mucosal-based gene signature and microbial shift for both tissue injury and clinical severity. We were able to characterize robust gene expression differences associated with deep ulcers and enriched pathway associations with oxidative stress and Th1 polarization. These data can serve to direct future therapy for ileal mucosal healing, which may not be achieved with current approaches (12). In contrast, we were able to detect an association between depletion of specific Firmicutes and Bacteroidetes taxa and expansion of Proteobacteria and clinical severity, as measured by the PCDAI. This would suggest that modification of the ileal microbial community may be required to achieve sustained clinical remission. This concept is supported by our regression model that included baseline ileal *APOA1* expression and specific microbiota, which was more accurate than one based only on clinical factors and anti–TNF-α exposure in predicting 6-month SSFR. Taken together, these data suggest that, in order to have a more durable effect, future therapeutic approaches for CD will need to address these specific microbial shifts, oxidative antimicrobial responses, and the profound loss of basal nuclear receptor-dependent homeostatic mechanisms.

Methods

More comprehensive information can be found in the Supplemental Methods.

The RISK cohort. Ileal biopsy samples and associated clinical information were obtained from the RISK study, an ongoing, prospective observational IBD inception cohort sponsored by the Crohn's and Colitis Foundation of America. 1,656 children and adolescents younger than 17 years, newly diagnosed with IBD and non-IBD Ctls, were enrolled at 28 North American pediatric gastroenterology centers between 2008 and 2012. All patients were required to undergo baseline colonoscopy and confirmation of characteristic chronic active colitis/ileitis by histology prior to diagnosis and treatment, with the recording of findings in standardized fashion. Only subjects with a confirmed persisting diagnosis of CD, UC, or Ctl during an average of 22 months follow-up to date were included in this analysis, which included a representative subgroup of age-matched CD (n = 243), Ctl (n = 43), and disease Ctl UC (n = 73) patients.

Ileal DNA and RNA extraction and RNA-seq. Ileal biopsies were obtained at the diagnostic colonoscopy and stored in RNALater at –80°C. Total DNA and RNA were isolated using the Qiagen AllPrep RNA/DNA Mini Kit. PolyA-RNA selection, fragmentation, cDNA synthesis, adaptor ligation, and library preparation were performed using TruSeq RNA Sample Preparation (Illumina). Single-end 50-bp sequencing was performed using the Illumina HiSeq 2000 in the CCHMC NIH-supported Digestive Health Center. Reads were aligned using TopHat (40). The aligned reads were quantified by Avadis NGS software (version 1.3.0, build 163982, Strand Scientific Intelligence Inc.) using Hg19 as the reference genome and reads per kilobase per million mapped reads (RPKM) as an output. The DESeq algorithm was used for RPKM normalization within Avadis NGS software. Only 12,415 transcripts with RPKM above 5 in 5 different samples were included in our downstream differential expression analysis.

RNA-seq expression and gene enrichment analysis. Samples were stratified into specific clinical subgroups, including Ctl, UC, cCD, iCD, iCD-noDU, and iCD-DU. For some analyses, cCD were further subdivided into those with and without microscopic inflammation on ileal biopsies. Differentially expressed genes were determined by the Audic Claverie method using the Benjamini-Hochberg false discovery rate correction (FDR correction 0.05) and analyzed for fold change differences as indicated. Normalized intensity values were used for unsupervised hierarchical clustering using Euclidean distance metric and Ward's linkage rule to test for groups of ileal biopsies with similar patterns of gene expression. Pearson correlation based on trend and rate of change was performed for *DUOX2* and *APOA1* gene expression as indicated across Ctl, UC, cCD, iCD-noDU, and iCD-DU for correlation coefficients of $0.98<|r|<1$. ToppGene (20), ToppCluster (21), and IPA (Ingenuity Systems) software were used to test for functional annotation enrichment analyses of upstream regulators, immune cell types, pathways, phenotype, and biologic functions. Functional annotation enrichment analyses for immune cell–type enrichments were

characterized using the Immunological Genome Project data series through ToppGene. Visualization of the network was obtained using Cytoscape.v3.0.2 (27).

Immunohistochemistry. Immunohistochemistry detection of APOA1, DUOX2, and a lipid peroxidation marker (4-hydroxy-2-nonenal [4-HNE]) was performed as previously described (41) using anti-APOA1 (Abcam, Ab75922), anti-DUOX2 (Santa Cruz Biotechnology, SC-49938), and 4-HNE (Bioss USA Antibodies, bs-6313R). Staining was examined using an Olympus BX51 light microscope and digitally recorded at ×40 magnification.

Support Vector Machine classification model to predict UC or cCD based on ileal gene expression. A Support Vector Machine–supervised classification algorithm included in Avadis was used to build a classification model for cCD and UC, using the cCD versus UC ileal gene expression signature (93 genes with fold change of 2.5) in the training cohorts (cCD1 and UC1). We then tested the accuracy of the model on the independent validation cohort (26 cCD2 and 28 UC2). We used the Avadis linear Support Vector Machine algorithm to build our prediction model on the training cohort (cCD1 and UC1) with its default parameters (maximum number of iteration = 100,000, cost = 100, ratio = 1, Kernel paraneter1 = 0.1, Kernel paraneter2 = 1, exponent = 2, sigma = 1). Building the model also included a 10-time cross-validation process using *N*-fold (*N* = 3), in which the classes in the input data are randomly divided into *N* equal parts; *N* – 1 parts are used for training, and the remaining 1 part is used for testing. Thus, each row is used at least once in training and once in testing, and a Confusion Matrix is generated. This model was then tested on the independent validation cohort (cCD2 and UC2).

Microbial community profiling and analysis of associations between microbial taxa and clinical and molecular metadata. Detailed protocols used for 16S amplification and sequencing are as previously described (42). In brief, 16S rRNA gene sequencing of ileal biopsy DNA was performed using the Illumina MiSeq v2 platform, targeting the V4 region of the SSU rRNA gene (primers: F [GTGCCAGCMGCCGCGGTAA] and R [GGACTACHVGGGTWTCTAAT]), according to the manufacturer's specifications with addition of 5% PhiX, and generating paired-end reads of 175 bp in length in each direction. The overlapping paired-end reads were stitched together, size selected, and further processed in a data curation pipeline implemented in QIIME (Quantitative Insights In to Microbial Ecology) 1.5.0 as pick_reference_otus.py (43). Taxonomy was assigned using the GreenGenes predefined taxonomy map of reference sequence operational taxonomic units (OTUs) to taxonomy (44) (version of May 2013). The resulting OTU tables were checked for mislabeling (45) and contamination (46) and further microbial community analysis and visualizations. A median sequence depth of 10,000 sequences per sample was obtained, and samples with less than 1,000 filtered sequences were excluded from analysis. OTUs were subsequently converted using QIIME to relative bacterial abundance. QIIME output was then trimmed down to the species level.

Multivariate analysis (http://huttenhower.sph.harvard.edu/galaxy) was performed as previously described (5, 6). Tests for association between taxa of the ileal microbial community and specific clinical and molecular metadata were conducted using MaAsLin. The following metadata were investigated in the analysis: clinical phenotype (Ctl, UC, CD), endoscopic severity (deep ulcers in ileum), clinical severity (PCDAI), and ileal gene expression of *APOA1*, *CXCL9*, *DUOXA2*, *LCT*, and *MUC4*. We controlled for age, gender, BMI (as a measure of nutritional status), and *NOD2*, *FUT2*, and *ATG16L1* IBD risk allele carriage in the analysis. Significant association was considered below a *q* value threshold of 0.25.

A biplot based on NMDS was used to visualize the relationship between the clinical and molecular metadata and the microbial taxa. The biplot uses points to represent samples, labels to represent selected significant microbial features, and labeled arrows to represent study metadata. Sample and microbial feature coordinates are generated as a standard biplot, with an additional dimension of metadata. Coordinates of metadata (Figure 6A, arrows) are determined by the center/average of the coordinates of the samples, with that metadata showing a central tendency of where that metadata is located. Stress is shown for the full ordination (both axes in Figure 6A) and can be interpreted as the percentage difference between current ordination and the data set in higher dimensions (ranging between no differences at 0.0 to complete difference at 1.0). Axes in Figure 6A represent the higher dimensional data set in 2 dimensions as approximated by NMDS. The full list of significant associations supporting the biplot is shown in Supplemental Excel file 16.

Regression analysis for 6-month SSFR. We used multiple logistic regression to account for the prognostic power of clinical and medication information and assess additional prognostic power resulting from including gene expression and microbial data in predicting SSFR after diagnosis. Clinical and medication information included in the models were age at diagnosis, baseline clinical severity defined by PCDAI (≤30 or >30), baseline mucosal severity defined by ileal deep ulceration (present or absent), and late anti-TNF therapy treatment (received or not). We excluded 7 patients with CD who received anti-TNF as initial therapy. Among the remaining 165 patients with CD, 27 received anti-TNF subsequent to other therapies (late anti-TNF therapy) prior to month 6. We considered 2 gene expression variables (*APOA1* and *DUOX2*) and microbial variables that were preidentified by the previous multivariate gene expression and microbiome analyses (Figure 6A and Supplemental Excel file 16). We then used variable selection and classification and regression tree analysis to construct 3 logistic regression models that respectively include clinical information only, clinical and significant gene expression variables, and clinical and significant gene expression and microbial variables. The reliability of the final model was tested by 10-fold cross-validation.

Statistics. Differentially expressed genes between groups were determined by the Audic Claverie method using the Benjamini-Hochberg FDR correction (FDR 0.05) and analyzed for fold change differences as indicated. Differences in expression for specific genes between patient subgroups were tested using Kruskal-Wallis with Dunn's multiple comparison test. $P < 0.05$ was considered significant. ToppGene (20) and ToppCluster (21) software was used to test for functional annotation enrichment of immune cell types, pathways, phenotype, and biologic functions with FDR $P < 0.05$. The Avadis linear Support Vector Machine–supervised classification algorithm was used with its default parameters (maximum number of iteration = 100,000, cost = 100, ratio = 1, Kernel paraneter1 = 0.1, Kernel paraneter2 = 1, exponent = 2, sigma = 1) to build a classification model for cCD and UC. Building the model also included a 10-time cross-validation process using *N*-fold (*N* = 3), in which the classes in the input data are randomly divided into *N* equal parts; *N* – 1 parts are used for training, and the remaining 1 part is used for testing. Multivariate analysis (http://huttenhower.sph.harvard.edu/galaxy) was performed as previously described (5, 6).

Significant association was considered below a *q* value threshold of 0.25. A biplot based on NMDS was used to visualize the relationship between the clinical and molecular metadata and the microbial taxa. The stress measurement tests for the goodness of fit of the biplot 2-dimensional depiction of the multidimensional data, with stress < 0.2 regarded as a good fit. For 6-month SSFR, we used multiple logistic regression to account for the prognostic power of clinical and medication information and assess additional prognostic power resulting from including gene expression and microbial data in predicting SSFR after diagnosis.

The RNA-seq data discussed in this publication have been deposited in NCBI's Gene Expression Omnibus (47) and are accessible with GEO series accession number GSE57945 (http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE57945). The microbial data were previously deposited as described (6).

Study approval. This study was approved by the institutional review boards at each of the participating RISK sites (Hospital for Sick Children; Cincinnati Children's Hospital Medical Center; Cedars-Sinai Medical Center, Los Angeles, California, USA; Cohen Children's Medical Center of New York; Children's Hospital of Philadelphia; Nationwide Children's Hospital; Goryeb Children's Hospital; Riley Children's Hospital, Indianapolis, Indiana, USA; IWK Health Centre, Halifax, Nova Scotia, Canada; UCSF; Hasbro Children's Hospital, Providence, Rhode Island, USA; Women & Children's Hospital of Buffalo, Buffalo, New York, USA; University of Utah and Primary Children's Medical Center, Salt Lake City, Utah, USA; Nemours Children's Clinic, Jacksonville, Florida, USA; Children's Hospital of Los Angeles, Los Angeles, California, USA; Baylor College of Medicine, Houston, Texas, USA; Children's Hospital of Wisconsin, Milwaukee, Wisconsin, USA; Children's Hospital of Eastern Ontario; Johns Hopkins Medical Center, Baltimore, Maryland, USA; University of Texas Southwestern Medical Center, Dallas, Texas, USA; University of Chicago, Chicago, Illinois, USA; Children's Hospital at Vanderbilt, Nashville, Tennessee, USA; Children's Healthcare of Atlanta, Atlanta, Georgia, USA; University of North Carolina at Chapel Hill, Chapel Hill, North Carolina, USA; Boston Children's Hospital, Boston, Massachusetts, USA; Emory University; Connecticut Children's Medical Center; Children's Hospital of Pittsburgh).

Acknowledgments

This work was supported by the Crohn's and Colitis Foundation of America, the Gene and Protein Expression and Bioinformatics cores of the NIH-supported Cincinnati Children's Hospital Research Foundation Digestive Health Center (1P30DK078392-01), the Cincinnati Children's Hospital Medical Center Innovation Fund (to L.A. Denson), NIH grant U54DK102557 (to R.J. Xavier, C. Huttenhower, and D. Gevers), and the Leona M. and Harry B. Helmsley Charitable Trust. We thank E. Bonkowski, B. Fey, and R. Steiner for excellent technical work. We thank the Crohn's and Colitis Foundation of America RISK study publication committee for critical review of this manuscript. We would also like to thank the following RISK study investigators: S.B. Snapper, R. Kellermayer, M. Kappelman, A. Otley, M. Pfefferkorn, S.A. Cohen, S.L. Guthery, N.S. LeLeiko, M. Oliva-Hemker, D.E. Moulton, B.S. Kirschner, A.S. Patel, and D.A. Ziring.

Address correspondence to: Lee A. Denson, Division of Pediatric Gastroenterology, Hepatology and Nutrition, Cincinnati Children's Hospital Medical Center, MLC 2010, 3333 Burnet Avenue, Cincinnati, Ohio 45229, USA. Phone: 513.636.7575; E-mail: lee.denson@cchmc.org.

1. Aujnarain A, Mack DR, Benchimol EI. The role of the environment in the development of pediatric inflammatory bowel disease. *Curr Gastroenterol Rep*. 2013;15(6):326.
2. Molodecky NA, et al. Increasing incidence and prevalence of the inflammatory bowel diseases with time, based on systematic review. *Gastroenterology*. 2012;142(1):46–54.e42
3. Kim SC, Tonkonogy SL, Karrasch T, Jobin C, Sartor RB. Dual-association of gnotobiotic IL-10$^{-/-}$ mice with 2 nonpathogenic commensal bacteria induces aggressive pancolitis. *Inflamm Bowel Dis*. 2007;13(12):1457–1466.
4. Jostins L, et al. Host-microbe interactions have shaped the genetic architecture of inflammatory bowel disease. *Nature*. 2012;491(7422):119–124.
5. Morgan XC, et al. Dysfunction of the intestinal microbiome in inflammatory bowel disease and treatment. *Genome Biol*. 2012;13(9):R79.
6. Gevers D, et al. The treatment-naive microbiome in new-onset Crohn's disease. *Cell Host Microbe*. 2014;15(3):382–392.
7. Hotte NS, et al. Patients with inflammatory bowel disease exhibit dysregulated responses to microbial DNA. *PLoS One*. 2012;7(5):e37932.
8. Colombel JF, et al. Infliximab, azathioprine, or combination therapy for Crohn's disease. *N Engl J Med*. 2010;362(15):1383–1395.
9. Sandborn WJ, et al. Ustekinumab induction and maintenance therapy in refractory Crohn's disease. *N Engl J Med*. 2012;367(16):1519–1528.
10. Parikh A, et al. Vedolizumab for the treatment of active ulcerative colitis: a randomized controlled phase 2 dose-ranging study. *Inflamm Bowel Dis*. 2012;18(8):1470–1479.
11. Walters TD, et al. Increased effectiveness of early therapy with anti-tumor necrosis factor-α vs an immunomodulator in children with Crohn's disease. *Gastroenterology*. 2014;146(2):383–391.
12. Arijs I, et al. Predictive value of epithelial gene expression profiles for response to infliximab in Crohn's disease. *Inflamm Bowel Dis*. 2010;16(12):2090–2098.
13. Gullberg E, Soderholm JD. Peyer's patches and M cells as potential sites of the inflammatory onset in Crohn's disease. *Ann N Y Acad Sci*. 2006;1072:218–232.
14. Jung C, Hugot JP, Barreau F. Peyer's patches: the immune sensors of the intestine. *Int J Inflam*. 2010;2010:823710.
15. Larsson E, et al. Analysis of gut microbial regulation of host gene expression along the length of the gut and regulation of gut microbial ecology through MyD88. *Gut*. 2012;61(8):1124–1131.
16. Zhang T, et al. An ileal Crohn's disease gene signature based on whole human genome expression profiles of disease unaffected ileal mucosal biopsies. *PLoS One*. 2012;7(5):e37139.
17. Zhang T, et al. Host genes related to paneth cells and xenobiotic metabolism are associated with shifts in human ileum-associated microbial composition. *PLoS One*. 2012;7(6):e30044.
18. Clark PM, Dawany N, Dampier W, Byers SW, Pestell RG, Tozeren A. Bioinformatics analysis reveals transcriptome and microRNA signatures and drug repositioning targets for IBD and other autoimmune diseases. *Inflamm Bowel Dis*. 2012;18(12):2315–2333.
19. Noble CL, et al. Characterization of intestinal gene expression profiles in Crohn's disease by genome-wide microarray analysis. *Inflamm Bowel Dis*. 2010;16(10):1717–1728.
20. Chen J, Bardes EE, Aronow BJ, Jegga AG. ToppGene Suite for gene list enrichment analysis and candidate gene prioritization. *Nucleic Acids Res*. 2009;37(Web Server issue):W305–W311.
21. Kaimal V, Bardes EE, Tabar SC, Jegga AG, Aronow BJ. ToppCluster: a multiple gene list feature analyzer for comparative enrichment clustering and network-based dissection of biological systems. *Nucleic Acids Res*. 2010; 38(Web Server issue):W96–W102.
22. Schwartz S, et al. A metagenomic study of diet-dependent interaction between gut microbiota and host in infants reveals differences in immune response. *Genome Biol*. 2012;13(4):r32.
23. Lipinski S, et al. DUOX2-derived reactive oxygen species are effectors of NOD2-mediated antibacterial responses. *J Cell Sci*. 2009; 122(pt 19):3522–3530.

24. Esworthy RS, Kim BW, Rivas GE, Leto TL, Doroshow JH, Chu FF. Analysis of candidate colitis genes in the Gdac1 locus of mice deficient in glutathione peroxidase-1 and -2. *PLoS One*. 2012;7(9):e44262.
25. Levy E, et al. Altered lipid profile, lipoprotein composition, and oxidant and antioxidant status in pediatric Crohn disease. *Am J Clin Nutr*. 2000;71(3):807–815.
26. Allez M, Lemann M, Bonnet J, Cattan P, Jian R, Modigliani R. Long term outcome of patients with active Crohn's disease exhibiting extensive and deep ulcerations at colonoscopy. *Am J Gastroenterol*. 2002;97(4):947–953.
27. Saito R, et al. A travel guide to Cytoscape plugins. *Nat Methods*. 2012;9(11):1069–1076.
28. Seeley EH, Washington MK, Caprioli RM, M'Koma AE. Proteomic patterns of colonic mucosal tissues delineate Crohn's colitis and ulcerative colitis. *Proteomics Clin Appl*. 2013; 7(7–8):541–549.
29. Schnitzler F, et al. Mucosal healing predicts long-term outcome of maintenance therapy with infliximab in Crohn's disease. *Inflamm Bowel Dis*. 2009;15(9):1295–1301.
30. Kuhn R, Lohler J, Rennick D, Rajewsky K, Muller W. Interleukin-10-deficient mice develop chronic enterocolitis. *Cell*. 1993;75(2):263–274.
31. Taurog JD, et al. The germfree state prevents development of gut and joint inflammatory disease in HLA-B27 transgenic rats. *J Exp Med*. 1994;180(6):2359–2364.
32. Frank DN, St Amand AL, Feldman RA, Boedeker EC, Harpaz N, Pace NR. Molecular-phylogenetic characterization of microbial community imbalances in human inflammatory bowel diseases. *Proc Natl Acad Sci U S A*. 2007;104(34):13780–13785.
33. Bornigen D, et al. Functional profiling of the gut microbiome in disease-associated inflammation. *Genome Med*. 2013;5(7):65.
34. Chappert P, Bouladoux N, Naik S, Schwartz RH. Specific gut commensal flora locally alters T cell tuning to endogenous ligands. *Immunity*. 2013;38(6):1198–1210.
35. Markle JG, et al. Sex differences in the gut microbiome drive hormone-dependent regulation of autoimmunity. *Science*. 2013; 339(6123):1084–1088.
36. Maslowski KM, Mackay CR. Diet, gut microbiota and immune responses. *Nat Immunol*. 2011;12(1):5–9.
37. Darsigny M, et al. Loss of hepatocyte-nuclear-factor-4α affects colonic ion transport and causes chronic inflammation resembling inflammatory bowel disease in mice. *PLoS One*. 2009;4(10):e7609.
38. Winter SE, et al. Host-derived nitrate boosts growth of E. coli in the inflamed gut. *Science*. 2013;339(6120):708–711.
39. Alexander EK, et al. Preoperative diagnosis of benign thyroid nodules with indeterminate cytology. *N Engl J Med*. 2012;367(8):705–715.
40. Trapnell C, Salzberg SL. How to map billions of short reads onto genomes. *Nat Biotechnol*. 2009;27(5):455–457.
41. Carey R, et al. Activation of an IL-6:STAT3-dependent transcriptome in pediatric-onset inflammatory bowel disease. *Inflamm Bowel Dis*. 2008;14(4):446–457.
42. Caporaso JG, et al. Ultra-high-throughput microbial community analysis on the Illumina HiSeq and MiSeq platforms. *ISME J*. 2012;6(8):1621–1624.
43. Caporaso JG, et al. QIIME allows analysis of high-throughput community sequencing data. *Nat Methods*. 2010;7(5):335–336.
44. McDonald D, et al. An improved Greengenes taxonomy with explicit ranks for ecological and evolutionary analyses of bacteria and archaea. *ISME J*. 2012;6(3):610–618.
45. Knights D, et al. Supervised classification of microbiota mitigates mislabeling errors. *ISME J*. 2011;5(4):570–573.
46. Knights D, et al. Bayesian community-wide culture-independent microbial source tracking. *Nat Methods*. 2011;8(9):761–763.
47. Edgar R, Domrachev M, Lash AE. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res*. 2002;30(1):207–210.