Supplementary Material

Patient Population and Clinical Phenotyping

Well-characterized CD patients from the adult IBD Center at University of North Carolina were included in this study (IRB Approval # 10-0355, 14-2445 and 11-0359). A total of 32 and 21 samples were submitted for RNA-seq and FAIRE-seq analyses, respectively. All non-genetic clinical phenotype data were collected by chart review and stored in a secured database. Clinical phenotypes included demographic and clinical variables: age, gender, disease duration, age at diagnosis, disease location, and type of disease behavior (Table 1; Supplementary Table 2). Mucosal biopsies were obtained from macroscopically unaffected sections of the ascending colon at time of surgery. These were also confirmed by an independent pathologist to have no active inflammation, only quiescent colitis. Tissue from non-IBD control patients was obtained at time for surgical resection for non-IBD related illness (Supplementary Table 3) and from a site distant from any pathology. The normal status of the area was confirmed by histology.

Tissue Isolation and Processing for RNA and DNA

Total RNA was isolated from flash-frozen tissue samples (mucosal not whole tissue) from surgical resections using the Qiagen RNeasy kit following the manufacturer's protocol. DNA for FAIRE was isolated from the same samples as previously described[1].

Cell Culture and transcriptional reporter assay

Human THP-1 acute monocytic leukemia cells were grown in RPMI-1640 supplemented with 10% fetal bovine serum and 1% penicillin/streptomycin at 37°C and 5% CO_2. THP-1 cells were seeded into 24-well plates (300,000 cells/well) 3 hours prior to transfection.

Regulatory elements were PCR-amplified from THP-1 genomic DNA using the primers below. Amplicons were cloned into the KpnI and XhoI restriction sites of the firefly luciferase reporter vector pGL4.23 in both the forward and reverse directions with respect to the minimal promoter. Individual clones were isolated and validated by Sanger sequencing. Each clone was transfected into THP-1 cells in duplicate (500 ng per well) using Lipofectamine 3000 and Opti-MEM. Additionally, phRL-TK *Renilla* luciferase reporter vector (40 ng) was transfected as an internal transfection efficiency control. After 48 hours, cell lysates were seeded into a 96-well plate in triplicate and luciferase activity was measured using the Dual Luciferase Reporter Assay System. Activity was normalized to the empty vector control. Two-tailed t-tests were performed on the raw firefly luciferase/renilla luciferase activity ratio.

SLC16A9	
Forward	TGATTAGTAGGCCTCTCTCTGT
Reverse	GCTCCTCTAGACTAGACTGATTG
SLC16A9 inverse	
Forward	GCTCCTCTAGACTAGACTGATTG
Reverse	TGATTAGTAGGCCTCTCTCTGT
DEPDC7	
Forward	AAGAGGTTAAATGATTTGCCCTG
Reverse	CCCATGCAATTGAAAATCCACA
DEPDC7 inverse	
Forward	CCCATGCAATTGAAAATCCACA
Reverse	AAGAGGTTAAATGATTTGCCCTG

RT-qPCR

Total RNA was isolated from flash-frozen tissue samples (mucosal not whole tissue) from surgical resections using the Qiagen RNeasy kit following the manufacturer's protocol. cDNA was derived from 1µg RNA by reverse transcriptase using the BioRad iScript cDNA Synthesis kit. RT-qPCR was then performed on these cDNA samples using the BioLine Hi-ROX SYBR kit with specific primers.

Gene	**Forward (5'-3')**	**Reverse (5'-3')**
PYGL	CACTCAAGTGGTCCTGGCTC	CGCATGGTGTTGACAGTGTT
PDK1	GGACTTCTACGCGCGCTTCT	AGCATTCACTGATCCGAAGTCC
CEACAM7	CACCCTGAATGTCCGCTATGA	CAGTCACTCTTCCCGAAATGC
APOA1	GCCTTGGGAAAACAGCTAAACC	CCAGAACTCCTGGGTCACA
SUSD2	CTCGGGACACTCAACAACGA	CATTGTGCACGGTCCAGTTG
XPNPEP	CACCCGTGTGCTGATAGGAA	CCACCATTCGCCCTGATGTA
GOLGA1	GAAACAGGACTTGGAGCAGC	ATGTTTGCCATCTCAGGTCC
GAPDH	CCAAGGTCATCCATGACAACTTTGGT	TGTTGAAGTCAGAGGAGACCACCTG

RNA isolation and RNA-seq analysis pipeline

Library preparation and mRNA sequencing were performed using protocols described previously[2]. In addition, all samples were genotyped using the Illumina Immunochip platform, and imputation was carried out with the MaCH-admix software[3]. Personalized genomes for each sample were created by incorporating known genetic variation from each individual. Paired-end 50-bp mRNA reads for each sample were then aligned to the corresponding personalized genome using GSNAP[4], with a kmer size of 15, two allowed mismatches per read, RefSeq splice site annotations, and the -v option for specifying heterozygous sites. A post-alignment blacklist step was used to filter reads that were aligned to problematic, highly-artefactual regions identified by ENCODE. This "allele-aware" alignment approach has been shown to greatly reduce mapping biases that arise due to discrepancies in genetic variation between an individual and the reference genome[5], and leads to a more accurate read count quantification.

Post-alignment quantification of RPKM values was conducted using an in-house script with RefSeq gene annotations, yielding a full set of 23,679 genes. Of these, a total of 14,873 genes were retained for differential expression analyses using the criteria that at least 10 samples had a RPKM > 1. Prior to analysis, RPKM values were incremented with a pseudocount of 1 and log normalized. Differentially expressed genes were called using DEseq[6] on raw counts for all genes, using an FDR cutoff of 0.05. Significance of overlap between previously published colon and ileum marker genes (947 genes) and differentially expressed genes between ileum-like and colon-like CD patients (849 genes) was determined using a hypergeometric test. A total of 106 genes overlapped between those up-regulated in colon-like patients (315 genes) and normal colon tissue (531 genes), and 183 genes overlapped between those up-regulated in ileum-like patients (534 genes) and normal ileum tissue (416 genes). Using 23,348 genes tested for differential expression as the population size, $P(X >= 106) = 2.67601007597266e\text{-}95$ for the colon-like overlap, and $P(X >= 183) = 4.090042811222961e\text{-}195$ for the ileum-like overlap.

Pediatric Crohn's disease expression data from ileal tissue was processed into RPKM as described previously[7], and downloaded from GEO (accession number GSE57945). A psuedocount of 1 and log normalization was applied to the pediatric

RPKM values, as described for the adult data. For joint analysis of pediatric ileal and adult colon expression data, we restricted to genes present in both data sets (22,525), and removed any genes highly expressed in one data set (mean RPKM > 5) but lowly expressed in the other (mean RPKM < 1), leading to a total of 21,881 genes. We then applied an additional quantile normalization step to the merged data matrix.

Principal components analysis

PCA analysis on adult individuals was performed using the log normalized RPKM values and the prcomp function in R. For the merged data matrix, the filtered, quantile-normalized set consisting of 21,881 genes was supplied to the prcomp function. For PCA analysis that included only pediatric individuals, prcomp was applied to the same set of 21,881 genes, using the log normalized RPKM values (prior to quantile normalization).

FAIRE and FAIRE-seq analysis pipeline

FAIRE was performed as described previously[1]. 50 bp single-end sequences were generated at UNC-CH HTSF using the Illumina HiSeq 2000 platform. Reads were filtered requiring a quality score of 20 or greater in at least 90 percent of nucleotides, and adapter contaminated reads were removed with TagDust[8]. Additionally, no more than 5 reads with identical sequence were retained. Non-filtered reads were aligned with SNP-tolerant GSNAP software[4] to personalized genomes, constructed as described above using k-mer size of 15 and allowing 1 mismatch per read. Post-alignment blacklist filtering was performed as described for RNA-seq reads.

The full genome was tiled into 300 bp windows overlapping by 100bp, and raw FAIRE-seq read overlaps were computed for each region. Windows overlapping with simple and low complexity repeat regions (as defined by RepeatMasker and downloaded from UCSC table browser) and the ENCODE DAC blacklist regions were masked from downstream analysis. Window counts were normalized by total aligned read counts for each sample, and batch effect correction was performed in R using ComBat[9].

Peaks were called using F-seq[10] with a feature size of 500 (-l option) and a user-supplied -bff background file for sequences of 50bp. A union set of peaks was created separately for each CD subclass (ileum-like and colon-like). A set of consistent peaks, defined as those peaks annotated in at least 30% of samples within a CD subclass, were created for each subclass. Peaks within 10bp were merged using the bedtools merge command with the –d 10 option, yielding a final union set of peaks for each CD subclass. To perform PCA, we first computed FAIRE signal at sliding 300-bp windows across the genome whose average normalized batch-corrected FAIRE signal was within a range of 10 to 100 and standard deviation exceeded 0.15. Resulting windows were then $\log_{10}$-transformed and median-centered._Differentially accessible regions (DARs) were identified using a two-sided t-test performed on normalized window counts for all 300 bp windows that intersected a peak in the consistent peak set for the two subclasses, as we have published previously[11]. When necessary, a single representative 300 bp window was selected from a group of overlapping 300 bp windows, where all were identified as a DAR, by selecting the 300 bp window with highest overall signal.

To compute the enrichment of DARs near differentially expressed genes, first the number of DARs falling within 50kb of a differentially expressed gene was recorded. To determine significance, we randomly created 1,000 sets of 300 bp windows taken from the consistent peaks for each subclass, where each random set consisted of the same number of windows as the computed set of DARs. For each of the 1,000 permutation sets, the number of windows falling within 50kb of a differentially expressed gene was

recorded. Significance was calculated empirically, by determining the number of permutation sets with a co-localization rate that exceeded the observed rate among the true DARs.

GWAS loci enrichment permutation

SNPs significantly associated with CD were downloaded from the NHGRI GWAS catalogue. In addition to the 163 tag SNPs represented in the catalogue, we included any SNP in high linkage disequillibrium ($r^2 > 0.8$), yielding a total of 3,179 disease-linked SNPs. We computed overlap between DARs and SNPs using bedtools [12]. For comparison with the observed overlap and to determine empirical significance, we created 1,000 sets of non-disease associated SNPs. Each null SNP set was created by first mapping the 163 tag SNPs to a randomly chosen, similarly-annotated non-disease linked SNP, based on high concordance of number of LD buddies ($r^2 > 0.8$), distance to nearest TSS, distance to nearest TES, and whether the SNP was in a gene and/or exon. The seed sets were subsequently expanded to include all highly linked SNPs ($r^2 > 0.8$), and were then overlapped with DARs using bedtools. The resulting overlap rates represent the null distribution expected under random chance, and was used to determine the statistical significance of the observed overlap statistic. P-values were computed by taking the number of times the overlap of a permuted set exceeded that of the observed set and multiplying by two to reflect a two-tailed distribution.

ChIP-seq analysis

Aligned ChIP-seq reads for histone marks H3K27ac, H3K27me3, H3K36me3, H3K4me1, H3K4me3, and H3K9me3 were downloaded from the Epigenome Roadmap project data portal[13], for both colonic mucosa and small intestinal tissue. Base pair resolution ChIP-seq signal was computed for 3 kb windows centered at the midpoint of DARs in the two CD subclasses. This signal was calculated by tallying the number aligned ChIP-seq reads overlapping each base pair for each DAR in a subclass, normalizing by sequencing depth of the ChIP-seq data set, and aggregating by the mean normalized read count across all DARs.

Selection of colon and ileum marker genes

For the pediatric samples, we selected the 50 pediatric ileum samples each that were most colon-like and most ileum-like based on the PCA (**Figure 2A**, second PC). Then, for each of the 947 genes previously identified to be differentially expressed between colon and ileum[14], we computed the standard deviation in normalized expression values across non-IBD, colon-like CD, and ileum-like CD samples in the adult and pediatric cohorts. We retained the top 500 most variably expressed genes and plotted their expression in these samples in **Figure 2B**.

Pathway Analysis

Pathway-level enrichments were calculated using GSAA (http://gsaa.unc.edu) [15, 16]. For RNA-seq data, genes were first ranked based on differential expression between samples from two classes (i.e. CD vs non-IBD) based on the t-statistics from the differential gene expression tests. The ranked list of t-statistics was input to GSAA, using the pre-ranked analysis option.

In order to create a chromatin-based score for each gene, we mapped all DARs within 100 kb of a gene's promoter to that gene, and then selected the most extreme t-statistic as the score for that gene. Genes not within 100 kb of any DAR were assigned scores of zero. In order to reduce cases of a gene being mapped to differential FAIRE-

seq signal in a nearby gene's promoter region (which is likely to represent a false positive), we took the additional step of masking all other gene's promoters when computing each gene's score. Gene scores were then input to GSAA as an ordered list, using the pre-ranked option.

Pathway enrichment was calculated for genesets derived from the Reactome Pathway Database (http://www.reactome.org/) [17]. Genesets with < 15 and > 500 genes were not considered. Genesets with an FDR <0.1 based on 2,000 permutations across all genes were considered significantly enriched.

To visualize pathways that were significantly enriched for a given contrast (FDR 5%), multidimensional scaling was used to find an optimal 2D arrangement of pathways based on a distance matrix of between-pathway semantic scores. Semantic scores were calculated using the *makeDendrogram.py* script of the python package pyEnrichment (https://github.com/ofedrigo/pyEnrichment, last accessed May 2016) and pathways were plotted using in-house R scripts.

Supplementary Figures

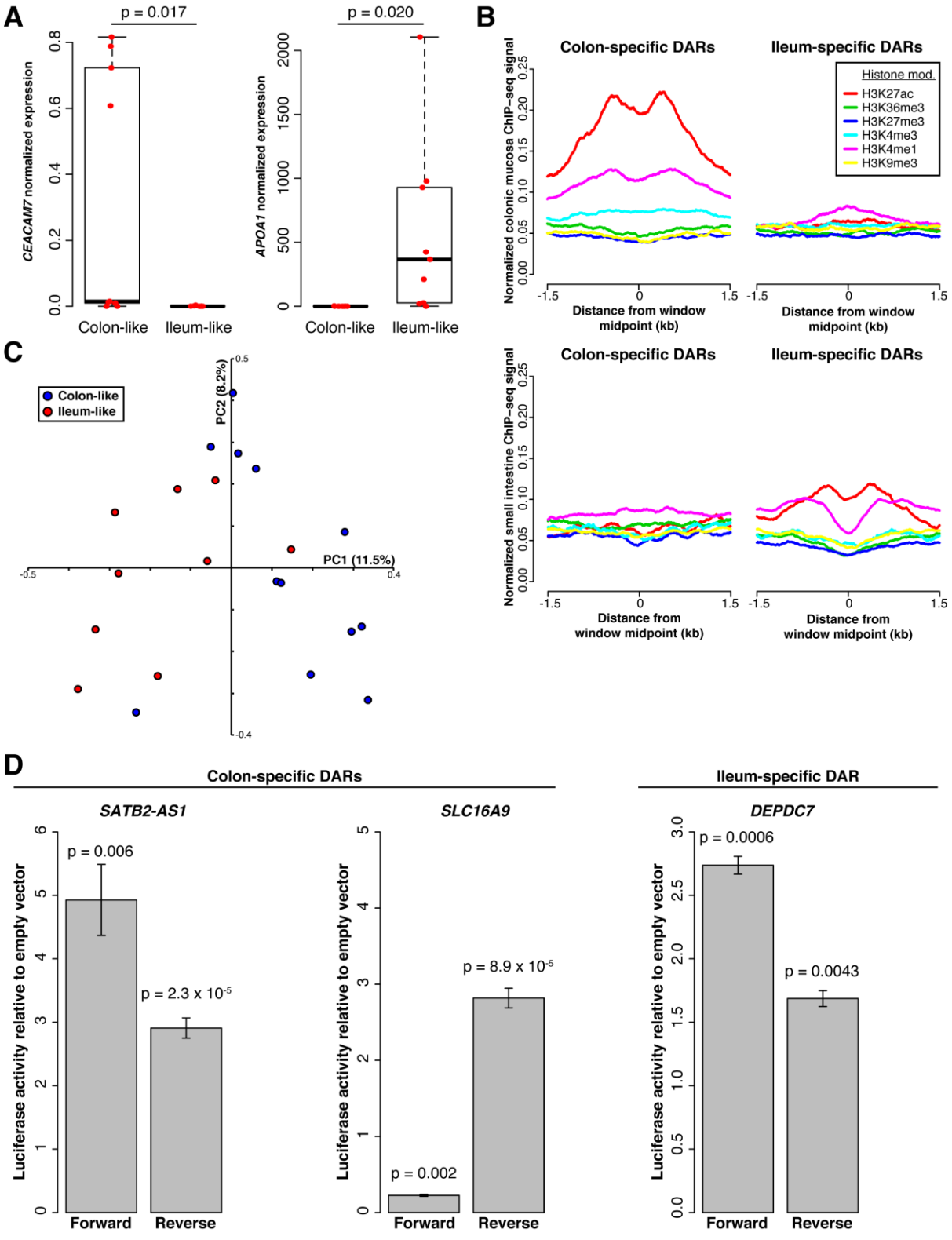
Supplementary Figure 1. Colon-like and ileum-like CD subclasses are demarcated by different tissue-specific histone modifications and enhancers and exhibit enhancer activity. A. qRT-PCR for *CEACAM7* (colon marker) and *APOA1* (ileum marker) expression in colon-like (n=11) and ileum-like (n=10) CD samples. P-values were computed using a two-sided t-test. **B.** Average normalized ChIP-seq enrichment for histone modifications detected in colonic mucosa (top) and small intestine (bottom) by the Epigenome Roadmap Consortium around colon-specific and ileum-specific regions of differential chromatin accessibility. **C.** Unsupervised principal components analysis (PCA) of normalized batch-corrected FAIRE-seq signal at sliding 300-bp windows with variable chromatin accessibility across samples. **D.** Luciferase activity normalized to empty vector controls for three regions of differential chromatin accessibility, cloned in both forward and reverse orientations.

Supplementary Figure 2. PCA analysis of pediatric CD patients reveals two disease subclasses in the ileum. Principal components 1 vs 2 (left) and 1 vs 3 (right) demonstrate that non-IBD and ileum-like CD samples group together, but that a separate group of colon-like samples segregate.

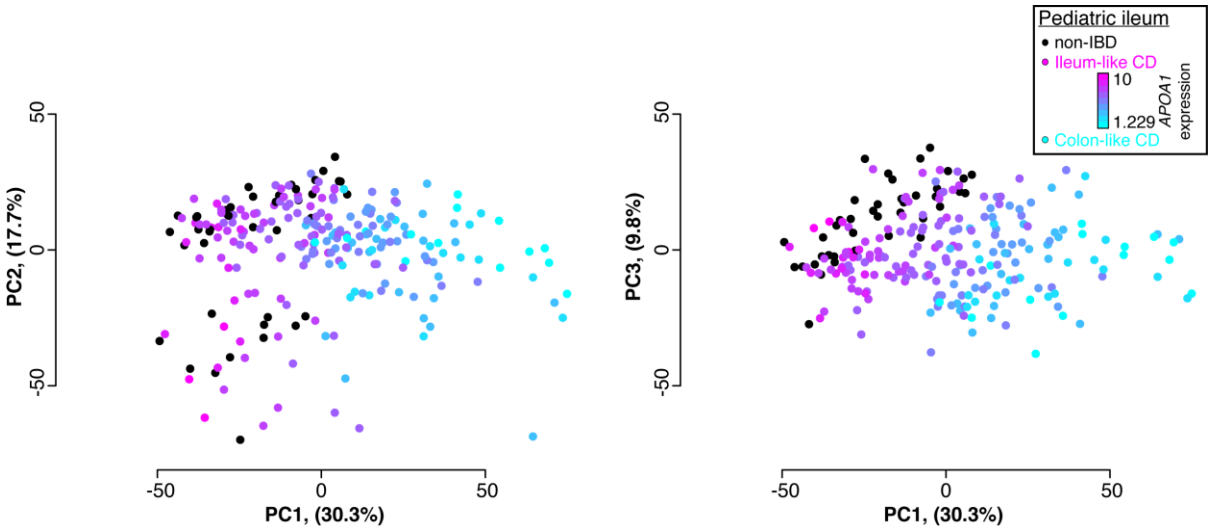
Supplementary Figure 3. Pathway-level enrichments for CD vs non-IBD controls in adult (colon tissue) and pediatric (ileum tissue) patients.

Supplementary Figure 4. Characterization and annotation of REACTOME pathways that were significantly differentially regulated between colon-like CD, ileum-like CD, and non-IBD samples from both adult and pediatric cohorts. Pathway sizes, categories based on direction of expression change, and broad annotation class labels are provided.

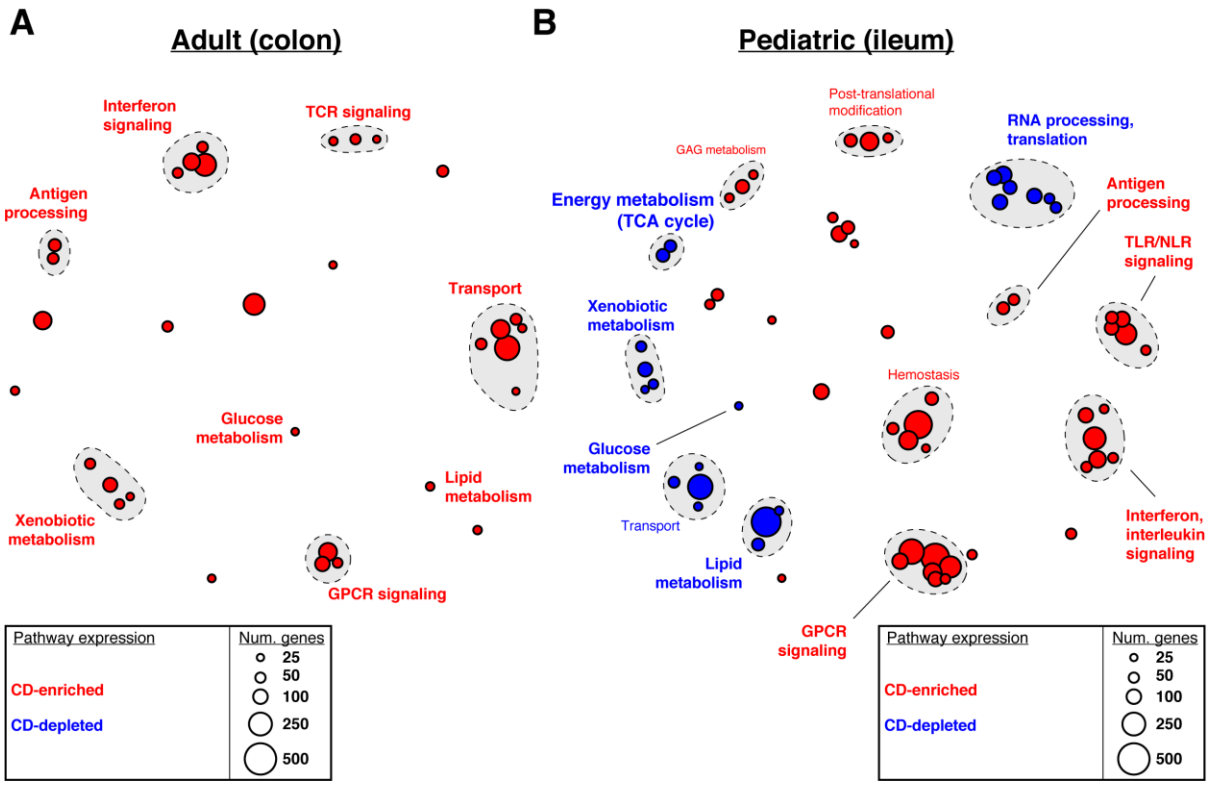
Supplement figure 1



Supplement figure 2



Supplement figure 3



Supplement figure 4

Pathway	Number of Genes	Adult Category CD vs nonIBD	Pediatric Category CD vs non-IBD	Adult Category Ileum-like vs Colon-like	Pediatric Category Ileum-like vs Colon-like	Broad Pathway Description
REACTOME_ANTIGEN_PROCESSING_CROSS_PRESENTATION	71	CD-enriched	CD-enriched		CL-enriched	Antigen Processing
REACTOME_ER_PHAGOSOME_PATHWAY	57	CD-enriched	CD-enriched			Antigen Processing
REACTOME_RESPIRATORY_ELECTRON_TRANSPORT_ATP_SYNTHESIS_BY_CHEMIOSMOTIC_COUPLING_AND_HEAT_PRODUCTION_BY_UNCOUPLING_PROTEINS_	79		CD-depleted	CL-enriched/IL-depleted	IL-enriched/CL-depleted	Energy Metabolism
REACTOME_RESPIRATORY_ELECTRON_TRANSPORT	65		CD-depleted	CL-enriched/IL-depleted	IL-enriched/CL-depleted	Energy Metabolism
REACTOME_TCA_CYCLE_AND_RESPIRATORY_ELECTRON_TRANSPORT	114			CL-enriched/IL-depleted	IL-enriched/CL-depleted	Energy Metabolism
REACTOME_PYRUVATE_METABOLISM_AND_CITRIC_ACID_TCA_CYCLE	39				IL-enriched/CL-depleted	Energy Metabolism
REACTOME_CHONDROITIN_SULFATE_DERMATAN_SULFATE_METABOLISM	39		CD-enriched		CL-enriched	GAG Metabolism
REACTOME_GLYCOSAMINOGLYCAN_METABOLISM	86		CD-enriched		CL-enriched	GAG Metabolism
REACTOME_HEPARAN_SULFATE_HEPARIN_HS_GAG_METABOLISM	37		CD-enriched			GAG Metabolism
REACTOME_GLUCONEOGENESIS	27	CD-enriched	CD-depleted		CL-depleted	Glucose Metabolism
REACTOME_GLUCOSE_METABOLISM	55				IL-enriched/CL-depleted	Glucose Metabolism
REACTOME_CHEMOKINE_RECEPTORS_BIND_CHEMOKINES	46	CD-enriched	CD-enriched	IL-enriched	CL-enriched	GPCR Signaling
REACTOME_PEPTIDE_LIGAND_BINDING_RECEPTORS	101	CD-enriched	CD-enriched	IL-enriched	CL-enriched	GPCR Signaling
REACTOME_CLASS_A1_RHODOPSIN_LIKE_RECEPTORS	154	CD-enriched	CD-enriched	IL-enriched	CL-enriched	GPCR Signaling
REACTOME_GPCR_LIGAND_BINDING	215		CD-enriched	IL-enriched	CL-enriched	GPCR Signaling
REACTOME_G_ALPHA_I_SIGNALLING_EVENTS	117		CD-enriched		CL-enriched	GPCR Signaling
REACTOME_SIGNALING_BY_GPCR	373		CD-enriched		CL-enriched	GPCR Signaling
REACTOME_GPCR_DOWNSTREAM_SIGNALING	291		CD-enriched		CL-enriched	GPCR Signaling
REACTOME_PLATELET_ACTIVATION_SIGNALING_AND_AGGREGATION	163		CD-enriched		CL-enriched	Hemostasis
REACTOME_RESPONSE_TO_ELEVATED_PLATELET_CYTOSOLIC_CA2_	63		CD-enriched		CL-enriched	Hemostasis
REACTOME_HEMOSTASIS	358		CD-enriched		CL-enriched	Hemostasis
REACTOME_GPVI_MEDIATED_ACTIVATION_CASCADE	31		CD-enriched		CL-enriched/IL-depleted	Hemostasis
REACTOME_INTERFERON_GAMMA_SIGNALING	54	CD-enriched	CD-enriched		CL-enriched	Interferon Signaling
REACTOME_INTERFERON_ALPHA_BETA_SIGNALING	48	CD-enriched	CD-enriched		CL-enriched	Interferon Signaling
REACTOME_INTERFERON_SIGNALING	134	CD-enriched	CD-enriched		CL-enriched	Interferon Signaling
REACTOME_CYTOKINE_SIGNALING_IN_IMMUNE_SYSTEM	234	CD-enriched	CD-enriched		CL-enriched	Interferon Signaling
REACTOME_SIGNALING_BY_ILS	101		CD-enriched		CL-enriched	Interleukin Signaling
REACTOME_IL1_SIGNALING	38		CD-enriched		CL-enriched	Interleukin Signaling
REACTOME_IL_3_5_AND_GM_CSF_SIGNALING	39				CL-enriched/IL-depleted	Interleukin Signaling
REACTOME_IL_2_SIGNALING	37				CL-enriched/IL-depleted	Interleukin Signaling
REACTOME_LIPID_DIGESTION_MOBILIZATION_AND_TRANSPORT	35	CD-enriched	CD-depleted	IL-enriched	CL-depleted	Lipid Metabolism
REACTOME_METABOLISM_OF_LIPIDS_AND_LIPOPROTEINS	403		CD-depleted	IL-enriched	CL-depleted	Lipid Metabolism
REACTOME_GLYCEROPHOSPHOLIPID_BIOSYNTHESIS	71		CD-depleted		CL-depleted	Lipid Metabolism
REACTOME_FATTY_ACID_TRIACYLGLYCEROL_AND_KETONE_BODY_METABOLISM	154				CL-depleted	Lipid Metabolism
REACTOME_PHOSPHOLIPID_METABOLISM	171				CL-depleted	Lipid Metabolism
REACTOME_INNATE_IMMUNE_SYSTEM	212	CD-enriched	CD-enriched		CL-enriched	NOD/TLR Signaling
REACTOME_TOLL_RECEPTOR_CASCADES	113		CD-enriched		CL-enriched	NOD/TLR Signaling
REACTOME_ACTIVATED_TLR4_SIGNALLING	90		CD-enriched		CL-enriched	NOD/TLR Signaling
REACTOME_NUCLEOTIDE_BINDING_DOMAIN_LEUCINE_RICH_REPEAT_CONTAINING_RECEPTOR_NLR_SIGNALING_PATHWAYS	44		CD-enriched		CL-enriched	NOD/TLR Signaling
REACTOME_NFKB_AND_MAP_KINASES_ACTIVATION_MEDIATED_BY_TLR4_SIGNALING_REPERTOIRE	70		CD-enriched		CL-enriched	NOD/TLR Signaling
REACTOME_MYD88_MAL_CASCADE_INITIATED_ON_PLASMA_MEMBRANE	80				CL-enriched	NOD/TLR Signaling
REACTOME_TRIF_MEDIATED_TLR3_SIGNALING	72				CL-enriched	NOD/TLR Signaling
REACTOME_NOD1_2_SIGNALING_PATHWAY	29				CL-enriched	NOD/TLR Signaling
REACTOME_O_LINKED_GLYCOSYLATION_OF_MUCINS	41		CD-enriched			Post-translational Modification
REACTOME_POST_TRANSLATIONAL_PROTEIN_MODIFICATION	155		CD-enriched			Post-translational Modification
REACTOME_ASPARAGINE_N_LINKED_GLYCOSYLATION	76		CD-enriched			Post-translational Modification
REACTOME_3_UTR_MEDIATED_TRANSLATIONAL_REGULATION	102		CD-depleted	IL-depleted	IL-depleted	RNA Processing, Translation
REACTOME_PEPTIDE_CHAIN_ELONGATION	83		CD-depleted	IL-depleted	IL-depleted	RNA Processing, Translation
REACTOME_INFLUENZA_LIFE_CYCLE	133		CD-depleted	IL-depleted	IL-depleted	RNA Processing, Translation
REACTOME_TRANSLATION	142			IL-depleted	IL-depleted	RNA Processing, Translation
REACTOME_METABOLISM_OF_RNA	251			IL-depleted	IL-depleted	RNA Processing, Translation
REACTOME_SRP_DEPENDENT_COTRANSLATIONAL_PROTEIN_TARGETING_TO_MEMBRANE	106			IL-depleted	IL-depleted	RNA Processing, Translation
REACTOME_INFLUENZA_VIRAL_RNA_TRANSCRIPTION_AND_REPLICATION	99		CD-depleted	IL-depleted		RNA Processing, Translation
REACTOME_NONSENSE_MEDIATED_DECAY_ENHANCED_BY_THE_EXON_JUNCTION_COMPLEX	104		CD-depleted	IL-depleted		RNA Processing, Translation
REACTOME_FORMATION_OF_THE_TERNARY_COMPLEX_AND_SUBSEQUENTLY_THE_43S_COMPLEX	48		CD-depleted	IL-depleted		RNA Processing, Translation
REACTOME_ACTIVATION_OF_THE_MRNA_UPON_BINDING_OF_THE_CAP_BINDING_COMPLEX_AND_EIFS_AND_SUBSEQUENT_BINDING_TO_43S	55		CD-depleted	IL-depleted		RNA Processing, Translation
REACTOME_METABOLISM_OF_NON_CODING_RNA	48				IL-depleted	RNA Processing, Translation
REACTOME_TRANSPORT_OF_MATURE_TRANSCRIPT_TO_CYTOPLASM	52				IL-depleted	RNA Processing, Translation
REACTOME_TRANSPORT_OF_MATURE_MRNA_DERIVED_FROM_AN_INTRONLESS_TRANSCRIPT	32				IL-depleted	RNA Processing, Translation
REACTOME_IMMUNOREGULATORY_INTERACTIONS_BETWEEN_A_LYMPHOID_AND_A_NON_LYMPHOID_CELL	52	CD-enriched	CD-enriched	IL-enriched	CL-enriched	TCR/BCR Signaling
REACTOME_COSTIMULATION_BY_THE_CD28_FAMILY	61	CD-enriched			CL-enriched	TCR/BCR Signaling
REACTOME_CD28_CO_STIMULATION	31				CL-enriched	TCR/BCR Signaling
REACTOME_ADAPTIVE_IMMUNE_SYSTEM	472				CL-enriched	TCR/BCR Signaling
REACTOME_TCR_SIGNALING	50	CD-enriched			CL-enriched/IL-depleted	TCR/BCR Signaling
REACTOME_GENERATION_OF_SECOND_MESSENGER_MOLECULES	26	CD-enriched			CL-enriched/IL-depleted	TCR/BCR Signaling
REACTOME_ANTIGEN_ACTIVATES_B_CELL_RECEPTOR_LEADING_TO_GENERATION_OF_SECOND_MESSENGERS	29				IL-depleted	TCR/BCR Signaling
REACTOME_SIGNALING_BY_THE_B_CELL_RECEPTOR_BCR	120				IL-depleted	TCR/BCR Signaling
REACTOME_DOWNSTREAM_TCR_SIGNALING	35	CD-enriched				TCR/BCR Signaling
REACTOME_TRANSPORT_OF_GLUCOSE_AND_OTHER_SUGARS_BILE_SALTS_AND_ORGANIC_ACIDS_METAL_IONS_AND_AMINE_COMPOUNDS	57	CD-enriched	CD-depleted	IL-enriched	CL-depleted	Transport
REACTOME_TRANSMEMBRANE_TRANSPORT_OF_SMALL_MOLECULES	286	CD-enriched	CD-depleted	IL-enriched	CL-depleted	Transport
REACTOME_ABC_FAMILY_PROTEINS_MEDIATED_TRANSPORT	25		CD-depleted	IL-enriched	CL-depleted	Transport
REACTOME_AMINO_ACID_AND_OLIGOPEPTIDE_SLC_TRANSPORTERS	33	CD-enriched	CD-depleted	IL-enriched		Transport
REACTOME_TRANSPORT_OF_INORGANIC_CATIONS_ANIONS_AND_AMINO_ACIDS_OLIGOPEPTIDES	59	CD-enriched		IL-enriched		Transport
REACTOME_SLC_MEDIATED_TRANSMEMBRANE_TRANSPORT	162	CD-enriched		IL-enriched		Transport
REACTOME_ABC_FAMILY_PROTEINS_MEDIATED_TRANSPORT	25	CD-enriched				Transport
REACTOME_MITOCHONDRIAL_PROTEIN_IMPORT	49			IL-depleted		Uncategorized
REACTOME_TRNA_AMINOACYLATION	42			IL-depleted		Uncategorized
REACTOME_ENDOSOMAL_SORTING_COMPLEX_REQUIRED_FOR_TRANSPORT_ESCRT	26				CL-depleted	Uncategorized
REACTOME_EXTRACELLULAR_MATRIX_ORGANIZATION	66		CD-enriched		CL-enriched	Uncategorized
REACTOME_INTEGRIN_CELL_SURFACE_INTERACTIONS	71		CD-enriched		CL-enriched	Uncategorized
REACTOME_UNFOLDED_PROTEIN_RESPONSE	77		CD-enriched		CL-enriched	Uncategorized
REACTOME_NCAM1_INTERACTIONS	28		CD-enriched		CL-enriched	Uncategorized
REACTOME_LATENT_INFECTION_OF_HOMO_SAPIENS_WITH_MYCOBACTERIUM_TUBERCULOSIS	26		CD-enriched		CL-enriched	Uncategorized
REACTOME_CELL_SURFACE_INTERACTIONS_AT_THE_VASCULAR_WALL	77		CD-enriched		CL-enriched	Uncategorized
REACTOME_PERK_REGULATED_GENE_EXPRESSION	26		CD-enriched		CL-enriched	Uncategorized
REACTOME_NCAM_SIGNALING_FOR_NEURITE_OUT_GROWTH	48				CL-enriched	Uncategorized
REACTOME_APC_C_CDC20_MEDIATED_DEGRADATION_OF_MITOTIC_PROTEINS	64				CL-enriched	Uncategorized
REACTOME_COLLAGEN_FORMATION	44		CD-enriched		CL-enriched/IL-depleted	Uncategorized
REACTOME_THE_ROLE_OF_NEF_IN_HIV1_REPLICATION_AND_DISEASE_PATHOGENESIS	27	CD-enriched			IL-depleted	Uncategorized
REACTOME_NEP_NS2_INTERACTS_WITH_THE_CELLULAR_EXPORT_MACHINERY	27				IL-depleted	Uncategorized
REACTOME_TRANSPORT_OF_RIBONUCLEOPROTEINS_INTO_THE_HOST_NUCLEUS	27				IL-depleted	Uncategorized
REACTOME_REGULATION_OF_GLUCOKINASE_BY_GLUCOKINASE_REGULATORY_PROTEIN	25				IL-depleted	Uncategorized
REACTOME_CELL_CYCLE_CHECKPOINTS	109				IL-depleted	Uncategorized
REACTOME_NUCLEAR_RECEPTOR_TRANSCRIPTION_PATHWAY	35	CD-enriched			IL-enriched/CL-depleted	Uncategorized
REACTOME_TRIGLYCERIDE_BIOSYNTHESIS	34	CD-enriched				Uncategorized
REACTOME_METABOLISM_OF_AMINO_ACIDS_AND_DERIVATIVES	149	CD-enriched				Uncategorized
REACTOME_TRANSPORT_TO_THE_GOLGI_AND_SUBSEQUENT_MODIFICATION	29	CD-enriched				Uncategorized
REACTOME_DIABETES_PATHWAYS	114		CD-enriched			Uncategorized
REACTOME_CLASS_B_2_SECRETIN_FAMILY_RECEPTORS	43		CD-enriched			Uncategorized
REACTOME_SIGNALING_BY_PDGF	110		CD-enriched			Uncategorized
REACTOME_ACTIVATION_OF_CHAPERONE_GENES_BY_XBP1S	46		CD-enriched			Uncategorized
REACTOME_PHASE1_FUNCTIONALIZATION_OF_COMPOUNDS	44	CD-enriched	CD-depleted	IL-enriched	CL-depleted	Xenobiotic Metabolism
REACTOME_CYTOCHROME_P450_ARRANGED_BY_SUBSTRATE_TYPE	28	CD-enriched	CD-depleted	IL-enriched	CL-depleted	Xenobiotic Metabolism
REACTOME_BIOLOGICAL_OXIDATIONS	95	CD-enriched	CD-depleted	IL-enriched	IL-enriched/CL-depleted	Xenobiotic Metabolism
REACTOME_PHASE_II_CONJUGATION	51	CD-enriched	CD-depleted	IL-enriched	IL-enriched/CL-depleted	Xenobiotic Metabolism

Supplementary Table 1

	Gene	baseMean	baseMeanIL	baseMeanCL	log2FoldChange	FDR
Colon-specific	NXPE4	794.7372	17.20361581	1501.58592	6.439365753	2.37E-49
	SATB2-AS1	129.6677	1.512208909	246.1727847	7.255074696	1.15E-41
	CWH43	284.0867	7.831195838	535.2280187	6.076740947	2.74E-41
	VSIG2	549.0849	22.56218639	1027.741953	5.503187579	3.35E-37
	CA2	7492.625	92.29712198	14220.19625	7.265887891	4.60E-22
	NXPE1	917.9078	128.8984973	1635.18905	3.664119501	6.38E-20
	B3GALT1	77.01145	2.857657527	144.4239827	5.610710097	1.03E-19
	GAL3ST2	44.51628	1.195573618	83.89874662	6.0187049	3.65E-17
	FOXD2	111.0708	17.71552037	195.9391437	3.459935222	1.37E-15
	CEACAM7	10201.14	77.75598868	19404.21291	7.961353618	2.90E-15
	LEFTY1	361.2523	9.816044775	680.7398663	6.10140692	2.57E-14
	SLPI	156.3061	25.78914762	274.9578363	3.409315634	3.44E-14
	AIFM3	374.2034	59.74500259	660.0746817	3.463545127	3.56E-14
	EYA2	135.8655	6.531586038	253.4417939	5.256725818	9.05E-13
	L1TD1	387.9812	19.28848066	723.1563512	5.221235428	1.56E-12
	ATP13A4	78.14697	13.87846262	136.5728826	3.28944942	2.05E-12
	C10orf99	952.5956	191.0196081	1644.937378	3.105572659	7.33E-12
	RHBDL2	111.8022	15.02319539	199.7831916	3.724322255	8.41E-12
	TFCP2L1	1317.31	134.4163809	2392.667042	4.15282617	1.77E-11
	PARM1	2306.327	550.1196899	3902.879409	2.826496178	3.11E-11
Ileum-specific	CPS1	5031.883	10519.44911	43.18748184	-7.924907277	7.92E-61
	CPS1-IT1	51.8355	108.5096878	0.313508808	-8.037019019	1.33E-36
	RBP2	3246.114	6785.64995	28.35453319	-7.897705947	8.57E-35
	CEACAM18	37.59547	78.36646426	0.530931261	-6.958449548	2.21E-29
	HTR1D	82.69207	169.0033704	4.227248286	-5.288311703	4.77E-29
	TM6SF2	426.1704	865.9806953	26.342828	-5.033553068	6.39E-26
	ABCC2	711.8181	1467.356445	24.96514509	-5.871491288	4.28E-25
	MTTP	7392.924	15471.31053	48.93696939	-8.301519199	4.28E-25
	ALPI	1527.674	3130.771863	70.31215346	-5.474596188	7.18E-24
	TMPRSS15	364.0587	763.2343771	1.171627182	-9.229495595	2.23E-23
	CCL25	1125.929	2356.507326	7.220973255	-8.330460217	2.31E-23
	SULT1E1	207.2398	427.6343316	6.881042843	-5.937128632	1.99E-22
	GSTA2	525.3816	1095.396045	7.186602982	-7.232122175	1.26E-21
	SLC2A2	1015.416	2127.691109	4.256362351	-8.932016851	3.65E-21
	CREB3L3	1546.829	3204.749616	39.62905799	-6.333918265	8.86E-20
	APOB	48064.94	100602.0138	303.9580369	-8.37009818	1.32E-19
	GSTA1	3152.422	6542.114402	70.88340128	-6.526153525	9.96E-18
	CLDN15	664.4674	1269.840222	114.1284281	-3.474766943	5.50E-17
	OTC	646.8354	1253.617774	95.21497145	-3.717365967	1.31E-16
	C19orf69	90.85682	190.2338293	0.514088476	-8.275869758	1.38E-16

Supplementary Table 1. Top 20 differentially expressed genes for each CD subclass. The normalized mean expression level within ileum-like (IL) and colon-like (CL) subclasses, as well as $\log_2$ fold-change and FDR values were generated by DEseq.

Supplementary Table 2

Phenotype	Colon-like											Ileum-like										
Patient ID	**51**	**54**	**62**	**63**	**405**	**407**	**408**	**420**	**429**	**431**	**440**	**20**	**21**	**25**	**29**	**64**	**413**	**422**	**424**	**434**	**450**	
Location																						
Ileum-only	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	1	1=yes; 0=no
Colon-only	0	1	0	0	1	1	1	1	0	1	0	0	0	0	0	0	1	0	0	0	0	1=yes; 0=no
Ileum+Colon	1	0	1	1	0	0	0	0	1	0	1	0	1	1	1	1	0	1	1	1	0	1=yes; 0=no
Patient Characteristics																						
Age at Surgery (years)	28	20	45	18	56	44	23	36	23	19	76	13	49	59	47	49	32	20	34	20	28	years
Male	0	1	0	1	0	0	0	0	0	0	1	1	0	0	0	0	0	0	1	1	0	1=yes; 0=no
Female	1	0	1	0	1	1	1	1	1	1	0	0	1	1	1	1	1	1	0	0	1	1=yes; 0=no
Smoker	1	1	1	0	0	0	0	0	0	1	0	0	0	1	1	0	1	1	1	0	0	1=current or previous smoker; 0=never smoked
Inflammation																						
Macroscopic	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1=yes; 0=no
Microscopic	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1=yes; 0=no
None	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1=yes; 0=no
Phenotypes and Involvement																						
Deep Ulcers	UNK	UNK	UNK	UNK	UNK	UNK	UNK	UNK	UNK	UNK	UNK	UNK	UNK	UNK	UNK	UNK	UNK	UNK	UNK	UNK	UNK	
Perianal	0	0	0	1	0	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1=yes; 0=no
Sigmoid	0	1	0	1	1	1	1	1	0	1	0	0	0	0	0	0	1	0	1	0	0	1=yes; 0=no
Rectal	1	1	0	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	1=yes; 0=no
Ileal Disease	1	0	1	1	0	0	0	0	1	0	1	1	1	1	1	1	0	1	1	1	1	1=yes; 0=no
Inflammatory	0	0	0	0	1	0	0	0	1	1	1	0	0	0	0	0	1	0	0	0	0	1=yes; 0=no
Stricturing	1	0	0	1	0	1	0	1	0	0	0	1	1	1	1	1	0	1	0	0	1	1=yes; 0=no
Penetrating	0	1	1	0	0	1	1	0	0	0	0	0	0	1	0	1	0	0	1	1	1	1=yes; 0=no
Pre-operative treament history																						
Steroids	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	0	1	1=yes; 0=no
5-ASA	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	0	1	1=yes; 0=no
Immunomodulation	0	0	0	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	0	1	1	1=yes; 0=no
Anti-TNF	1	0	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	0	0	1	1	1=yes; 0=no
Non-anti-TNF biologic	0	0	0	0	0	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	1=yes; 0=no
Post-operative outcome																						
Disease recurrence	0	UNK	0	0	0	0	0	0	0	0	0	UNK	0	1	UNK	0	0	0	1	0	0	1=yes; 0=no
Biologic Use	0	0	0	0	0	0	0	1	0	0	0	UNK	1	0	1	1	1	0	1	1	0	1=yes; 0=no
Colectomy	0	1	0	1	1	1	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	1=yes; 0=no
Second resection	0	0	1	1	1	1	0	0	1	0	1	0	0	1	1	0	0	0	0	1	0	1=yes; 0=no
Time to first resection (years)	6	5	6	6	1	19	5	23	4	3	4	2	9	0	14	33	13	4	2	0	2	years
Time from first to second resection (years, if applicable)	NA	NA	1	3	15	1	NA	NA	1	NA	47	NA	NA	3	5	NA	NA	NA	NA	1	NA	years

Supplemental Table 2: Charactersitics of individual adult Crohn's disease patients. NA = Not applicable. UNK = unknown.

Supplementary Table 2: Characteristics of individual adult Crohn's disease patients. NA = Not applicable. UNK = unknown.

Supplementary Table 3

Patient ID	**22**	**23**	**27**	**30**	**32**	**36**	**39**	**43**	**48**	**49**	**50**	
Location												
Ileum-only	0	0	0	0	0	0	0	0	0	0	0	1=yes; 0=no
Colon-only	1	1	1	1	1	1	1	1	1	1	1	1=yes; 0=no
Ileum+Colon	0	0	0	0	0	0	0	0	0	0	0	1=yes; 0=no
Patient Characteristics												
Age at Surgery (years)	53	52	70	82	44	41	52	45	70	62	49	years
Male	0	1	1	0	1	0	1	0	1	1	0	1=yes; 0=no
Female	1	0	0	1	0	1	0	1	0	0	1	1=yes; 0=no
Smoker												1=current or previous smoker; 0=never smoked
Inflammation												
Macroscopic	0	0	0	0	0	0	0	0	0	0	0	1=yes; 0=no
Microscopic	0	0	0	0	0	0	0	0	0	0	0	1=yes; 0=no
None	1	1	1	1	1	1	1	1	1	1	1	1=yes; 0=no
Disease Phenotypes												
Colon Cancer	1	1	0	1	1	0	1	0	0	0	0	
Diverticulitis	0	0	1	0	0	0	0	0	0	0	0	1=yes; 0=no
Colonic Inertia	0	0	0	0	0	1	0	1	0	0	0	1=yes; 0=no
Adenoma	0	0	0	0	0	0	0	0	1	1	0	1=yes; 0=no
Si Neuroendocrine Tumor	0	0	0	0	0	0	0	0	0	0	1	1=yes; 0=no
Pre-operative treament history												
Steroids												1=yes; 0=no
5-ASA												1=yes; 0=no
Immunomodulation												1=yes; 0=no
Anti-TNF												1=yes; 0=no
Non-anti-TNF biologic												1=yes; 0=no
Post-operative outcome												
Disease recurrence												1=yes; 0=no
Biologic Use												1=yes; 0=no
Colectomy												1=yes; 0=no
Second resection												1=yes; 0=no
Time to first resection (years)												years
Time from first to second resection (years, if applicable)												years

Supplemental Table 3: Charactersitics of individual adult non-IBD patients.

Supplementary Table 3: Characteristics of individual adult non-IBD patients.

REFERENCES

1 Simon JM, Giresi PG, Davis IJ, Lieb JD. Using formaldehyde-assisted isolation of regulatory elements (FAIRE) to isolate active regulatory DNA. Nature protocols 2012;**7**:256-67.

2 Peck BC, Weiser M, Lee SE, Gipson GR, Iyer VB, Sartor RB*, et al.* MicroRNAs Classify Different Disease Behavior Phenotypes of Crohn's Disease and May Have Prognostic Utility. Inflamm Bowel Dis 2015;**21**:2178-87.

3 Liu EY, Li M, Wang W, Li Y. MaCH-admix: genotype imputation for admixed populations. Genet Epidemiol 2013;**37**:25-37.

4 Wu TD, Nacu S. Fast and SNP-tolerant detection of complex variants and splicing in short reads. Bioinformatics 2010;**26**:873-81.

5 Buchkovich ML, Eklund K, Duan Q, Li Y, Mohlke KL, Furey TS. Removing reference mapping biases using limited or no genotype data identifies allelic differences in protein binding at disease-associated loci. BMC Med Genomics 2015;**8**:43.

6 Anders S, Huber W. Differential expression analysis for sequence count data. Genome biology 2010;**11**:R106.

7 Haberman Y, Tickle TL, Dexheimer PJ, Kim MO, Tang D, Karns R*, et al.* Pediatric Crohn disease patients exhibit specific ileal transcriptome and microbiome signature. J Clin Invest 2014;**124**:3617-33.

8 Lassmann T, Hayashizaki Y, Daub CO. TagDust-a program to eliminate artifacts from next generation sequencing data. Bioinformatics 2009;**25**:2839-40.

9 Johnson WE, Li C, Rabinovic A. Adjusting batch effects in microarray expression data using empirical Bayes methods. Biostatistics 2007;**8**:118-27.

10 Boyle AP, Guinney J, Crawford GE, Furey TS. F-Seq: a feature density estimator for high-throughput sequence tags. Bioinformatics 2008;**24**:2537-8.

11 Simon JM, Hacker KE, Singh D, Brannon AR, Parker JS, Weiser M*, et al.* Variation in chromatin accessibility in human kidney cancer links H3K36 methyltransferase loss with widespread RNA processing defects. Genome Research 2013;**xx**:yy-yy.

12 Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics 2010;**26**:841-2.

13 Roadmap Epigenomics C, Kundaje A, Meuleman W, Ernst J, Bilenky M, Yen A*, et al.* Integrative analysis of 111 reference human epigenomes. Nature 2015;**518**:317-30.

14 Comelli EM, Lariani S, Zwahlen MC, Fotopoulos G, Holzwarth JA, Cherbut C*, et al.* Biomarkers of human gastrointestinal tract regions. Mamm Genome 2009;**20**:516-27.

15 Xiong Q, Ancona N, Hauser ER, Mukherjee S, Furey TS. Integrating genetic and gene expression evidence into genome-wide association analysis of gene sets. Genome Res 2012;**22**:386-97.

16 Xiong Q, Mukherjee S, Furey TS. GSAASeqSP: a toolset for gene set association analysis of RNA-Seq data. Sci Rep 2014;**4**:6347.

17 Merico D, Isserlin R, Stueker O, Emili A, Bader GD. Enrichment map: a network-based method for gene-set enrichment visualization and interpretation. PLoS One 2010;**5**:e13984.