Comprehensive Supplemental Materials and Methods

The RISK cohort. Ileal biopsy samples and associated clinical information were obtained from the RISK study, an ongoing, prospective observational inflammatory bowel disease (IBD) inception cohort sponsored by the Crohn's and Colitis Foundation of America (CCFA). 1656 children and adolescents younger than 17 years newly diagnosed with IBD and non-IBD controls 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. Non-IBD controls were subjects suspected to have IBD, but with normal radiographic, endoscopic, and histologic findings. Once standard and published diagnostic guidelines were met, patients were either diagnosed with Crohn disease (CD), ulcerative colitis (UC), or IBD-U (IBD undefined) or were considered as non-IBD controls (Ctl). Only subjects with a confirmed 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 sub-group of age matched CD (n=243), Ctl (n=43) and disease control UC (n=73) patients (Supplemental Fig. 1).

Ethical considerations. This study was approved by the Institutional Review Boards at each of the participating RISK sites. All subjects and/or their parents/guardians provided informed consent, with pediatric subjects over the age of 11 providing assent.

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 according to the manufacturer’s instructions (QIAGEN, Valencia, CA). The quality and concentration of RNA was measured by the Agilent Bioanalyzer 2100 (Hewlett Packard) using the RNA 6000 Nano Assay to confirm a 28S/18S ratio of 1.6–2.0. Mean (95^{th}CI) yield of RNA and DNA (1-4 biopsies) was equal to 11,490(9,351-13,640) ng and 10,500(8468,12,670) ng per sample, respectively, with 90% having a RNA integrity number (RIN) > 7. PolyA-RNA selection, fragmentation, cDNA synthesis, adaptor ligation and library preparation was performed using TruSeqTM RNA Sample Preparation according to the manufacturer’s instructions (Illumina, San Diego, CA). Single end 50bp sequencing was performed using the Illumina HiSeq 2000 in the CCHMC NIH-supported Digestive Health Center Gene and Protein Expression core with mean(SD) coverage of 18,128,386(5,332,968) reads per sample. Reads were aligned using TopHat(1), which efficiently aligns reads spanning known or novel splice junctions. The aligned reads were quantified by Avadis® NGS software, (Version 1.3.0, Build 163982 ©Strand Scientific Intelligence, Inc., San Francisco, CA, USA.), 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 and the normalized counts were log2-transformed and base-lined to the median expression of control samples. Only 12,415 transcripts with RPKM above 5 in 5 different samples were included in our downstream differential expression analysis. Of note, the normalized signal values were already in log-scale.

RNA-seq expression and gene enrichment analysis. Samples were stratified into specific clinical sub-groups including Ctl, UC, colon-only CD (cCD), ileal CD (iCD), ileal CD without deep ulcers (iCD-noDU), and ileal CD with deep ulcers (iCD-DU). Differentially expressed genes were determined by the Audic Claverie method using the Benjamini–Hochberg false discovery rate correction (FDR, 0.05), and analyzed for fold change differences (FC) as indicated. Since the normalized values were already in a log scale, FC was computed as (direction of change) x $2^{|\log FC|}$, were log FC is the difference between the two selected conditions averages. Normalized intensity values (averaged or non-averaged as indicated) were used for hierarchical clustering of both genes expression and conditions using Euclidean distance metric and Ward's linkage rule. 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 co-efficients of $0.98<|r|<1$. ToppGene(2) and IPA (Ingenuity® Systems, www.ingenuity.com) software were used 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, reporting the top 80% annotation within each cell type category.

IPA(Ingenuity® Systems, www.ingenuity.com) top upstream regulator and the associated genes were selected for further analysis using ToppCluster (3) based on their activation z-score and associated p-values (Suppl. Table 9); PPPARG, HNF4A and STAT1 for the *APOA1* module and NR3C1 and NFKB complex for the *DUOX2* module. Each transcription factor was used as a node (cluster), where genes that were associated with the transcription factor based on the IPA analysis were used for pathway enrichments analyses (GO, Mouse phenotypes) with FDR correction (0.05) in ToppCluster to generate a network in Fig 1e. Only up to top 5 enrichments pathways were used to generate the network. Visualization of the network was further modified using Cytoscape.v3.0.2 (4).

Immunohistochemistry. Immunohistochemistry detection of APOA1, DUOX2, and a lipid peroxidation marker (4-hydroxy-2-nonenal (4-HNE)) was performed as previously described (5). 5µm paraffin-embedded slides were deparaffinized and antigen unmasking was carried out by boiling for 8 minutes with 10 mM sodium citrate (pH 6). Endogenous peroxide was quenched with 3% hydrogen peroxide for 15 minutes at RT, then tissues were permeabilized with 0.3%Triton in PBS for 15 minutes at RT. Slides were subsequently blocked with 3% normal serum for 60 minutes at RT, and then incubated overnight at 4°C with primary antibodies as follows: anti-APOA1 (Abcam, Cambridge, MA, Ab75922), anti-DUOX2 (Santa Cruz Biotechnology, Dallas, TX, SC-49938) and 4-hydroxy-2-nonenal (4-HNE) (Bioss USA Antibodies, Woburn, MA, bs-6313R). Biotinylated secondary antibody and avidin-biotin complex (Vector Laboratories, Burlingame, CA, PK-7100) were applied sequentially for 60 minutes at RT after washing in PBS. Hematoxylin was used for nuclear counterstaining following peroxidase (DAB substrate kit from Vector Laboratories, Burlingame, CA, SK-4100) development. After drying, slides were mounted using Permount (Fisher Scientific, Pittsburgh, PA, SP15-100). Staining was examined using an Olympus BX51 light microscope and digitally recorded at 40x 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, utilizing 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 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=100000, cost=100, ratio=1, Kernel paraneter1=0.1, Kernel paraneter2=1, exponent =2, sigma=1). Building the model also included a 10 times cross validation process using N-fold (N=3), where the classes in the input data are randomly divided into N equal parts; N-1 parts are used for training, and the remaining one 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 run 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 described before (6). 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 175b in length in each direction. The overlapping paired-end reads were stitched together (approximately 97 bp overlap), size selected to reduce non-specific amplification products from host DNA (225 - 275 bp), 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 (7). Taxonomy is assigned using the Greengenes predefined taxonomy map of reference sequence operational taxonomic units **(**OTUs) to taxonomy(8) (version of May 2013). The resulting OTUs tables are checked for mislabeling(9)and contamination(10), and further microbial community analysis and visualizations. A median sequence depth of 10,000/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, resulting in a final microbial output that contained 161 different microbial taxa.

Multivariate Analysis. Test for association between taxa of the ileal microbial community and specific clinical and molecular metadata were conducted using Multivariate Analysis by Linear Models (MaAsLin). The following metadata were investigated in the analysis: clinical phenotype (Ctl, UC, CD), endoscopic severity (deep ulcers in ileum), clinical severity (Pediatric Crohn Disease Activity Index, PCDAI), and ileal *APOA1, CXCL9*, *DUOXA2, LCT,* and *MUC4* gene expression. We controlled for age, gender, body mass index (BMI, as a measure of nutritional status), and *NOD2*, *FUT2*, and *ATG16L1* IBD risk allele carriage in the analysis. Samples included 180 CD, 36 UC, and 35 Ctl for whom RNASeq had been performed. A comprehensive description of this analysis method can be found online at http://huttenhower.sph.harvard.edu/galaxy and was previously described (11). In short, for each arcsine square-root transformed microbial feature, a model is selected from metadata using gradient boosting (gbm package(12)). Covariates in the selected model are then evaluated controlling for potential confounders using a general linear model. Within each metadatum/clade association independently, multiple comparisons over factor levels were adjusted using a Bonferonni correction; multiple hypothesis tests over all clades and metadata were adjusted to produce a final Benjamini and Hochberg false discovery rate(13). Significant association was considered below a q-value threshold of 0.25.

A biplot based on non-metric multidimensional scaling (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 (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. More specifically, discontinuous metadata are broken down to levels (values) and each level is made into its own binary metadata (0 for not having that value and 1 for having that value). For each discontinuous metadata level, samples with the value of 1 are selected and their coordinates in the ordination are averaged. This average coordinate set is then used as the coordinates for that metadata level. For continuous data, using the ordination coordinates for all the sample points, the value of the continuous metadata is placed in a landscape using the sample coordinates as x and y and the z as the metadata value. This is then smoothed with a lowess and then the maximum fitted value's coordinates are used as the coordinates of the central tendency of the metadata. Stress is shown for the full ordination (both axes) and can be interpreted as the percent 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 represent the higher dimensional data set in two dimensions as approximated by NMDS. The full list of significant associations supporting the biplot are shown in Suppl. Table 16.

Regression analysis for month six steroid- and surgery-free remission (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 steroid and surgery free remission six months 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 seven CD patients who received anti-TNF as initial therapy. Amongst the remaining 165 CD patients, 27 received anti-TNF subsequent to other therapies (late anti-TNF therapy) prior to month six. We considered two gene expression variables (*APOA1*, *DUOX2*) and statistical significant microbial variables that were pre-identified by the previous multivariate gene expression and microbiome analyses (Supplemental Table 16). We then used variable selection and classification and regression tree (CART) analysis to construct three logistic regression models that respectively include clinical information only, clinical and significant gene expression variables, and clinical and significant gene expression and microbial variables.

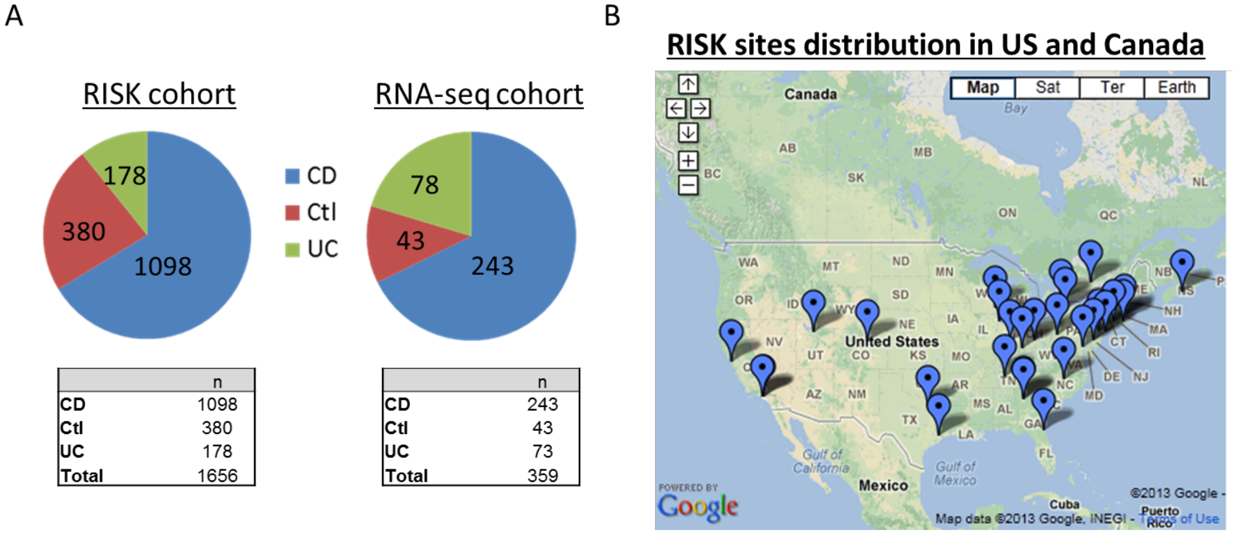
The RNA-seq data discussed in this publication have been deposited in NCBI's Gene Expression Omnibus (14) and are accessible through Geo Series accession number GSE57945 (http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE57945). The microbial data was previously deposited as described (15).

Methods References

1. Trapnell, C., and Salzberg, S.L. 2009. How to map billions of short reads onto genomes. *Nat Biotechnol* 27:455-457.
2. Chen, J., Bardes, E.E., Aronow, B.J., and Jegga, A.G. 2009. ToppGene Suite for gene list enrichment analysis and candidate gene prioritization. *Nucleic Acids Res* 37:W305-311.
3. Kaimal, V., Bardes, E.E., Tabar, S.C., Jegga, A.G., and Aronow, B.J. 2010. ToppCluster: a multiple gene list feature analyzer for comparative enrichment clustering and network-based dissection of biological systems. *Nucleic Acids Res* 38:W96-102.
4. Saito, R., Smoot, M.E., Ono, K., Ruscheinski, J., Wang, P.L., Lotia, S., Pico, A.R., Bader, G.D., and Ideker, T. 2012. A travel guide to Cytoscape plugins. *Nat Methods* 9:1069-1076.
5. Carey, R., Jurickova, I., Ballard, E., Bonkowski, E., Han, X., Xu, H., and Denson, L.A. 2008. Activation of an IL-6:STAT3-dependent transcriptome in pediatric-onset inflammatory bowel disease. *Inflamm Bowel Dis* 14:446-457.
6. Caporaso, J.G., Lauber, C.L., Walters, W.A., Berg-Lyons, D., Huntley, J., Fierer, N., Owens, S.M., Betley, J., Fraser, L., Bauer, M., et al. 2012. Ultra-high-throughput microbial community analysis on the Illumina HiSeq and MiSeq platforms. *ISME J* 6:1621-1624.
7. Caporaso, J.G., Kuczynski, J., Stombaugh, J., Bittinger, K., Bushman, F.D., Costello, E.K., Fierer, N., Pena, A.G., Goodrich, J.K., Gordon, J.I., et al. 2010. QIIME allows analysis of high-throughput community sequencing data. *Nat Methods* 7:335-336.

8. McDonald, D., Price, M.N., Goodrich, J., Nawrocki, E.P., DeSantis, T.Z., Probst, A., Andersen, G.L., Knight, R., and Hugenholtz, P. 2012. An improved Greengenes taxonomy with explicit ranks for ecological and evolutionary analyses of bacteria and archaea. *ISME J* 6:610-618.
9. Knights, D., Kuczynski, J., Koren, O., Ley, R.E., Field, D., Knight, R., DeSantis, T.Z., and Kelley, S.T. 2011. Supervised classification of microbiota mitigates mislabeling errors. *ISME J* 5:570-573.
10. Knights, D., Kuczynski, J., Charlson, E.S., Zaneveld, J., Mozer, M.C., Collman, R.G., Bushman, F.D., Knight, R., and Kelley, S.T. 2011. Bayesian community-wide culture-independent microbial source tracking. *Nat Methods* 8:761-763.
11. Morgan, X.C., Tickle, T.L., Sokol, H., Gevers, D., Devaney, K.L., Ward, D.V., Reyes, J.A., Shah, S.A., LeLeiko, N., Snapper, S.B., et al. 2012. Dysfunction of the intestinal microbiome in inflammatory bowel disease and treatment. *Genome Biol* 13:R79.
12. Friedman, J. 2001. Stochastic gradient boosting. *Computational Statistics & Data Analysis* 38:367-378.
13. Benjamini, Y., Hochberg Y. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Royal Statistical Society B* 57:289-300.
14. Edgar, R., Domrachev, M., and Lash, A.E. 2002. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res* 30:207-210.
15. Gevers, D., Kugathasan, S., Denson, L.A., Vazquez-Baeza, Y., Van Treuren, W., Ren, B., Schwager, E., Knights, D., Song, S.J., Yassour, M., et al. 2014. The treatment-naive microbiome in new-onset Crohn's disease. *Cell Host Microbe* 15:382-392.

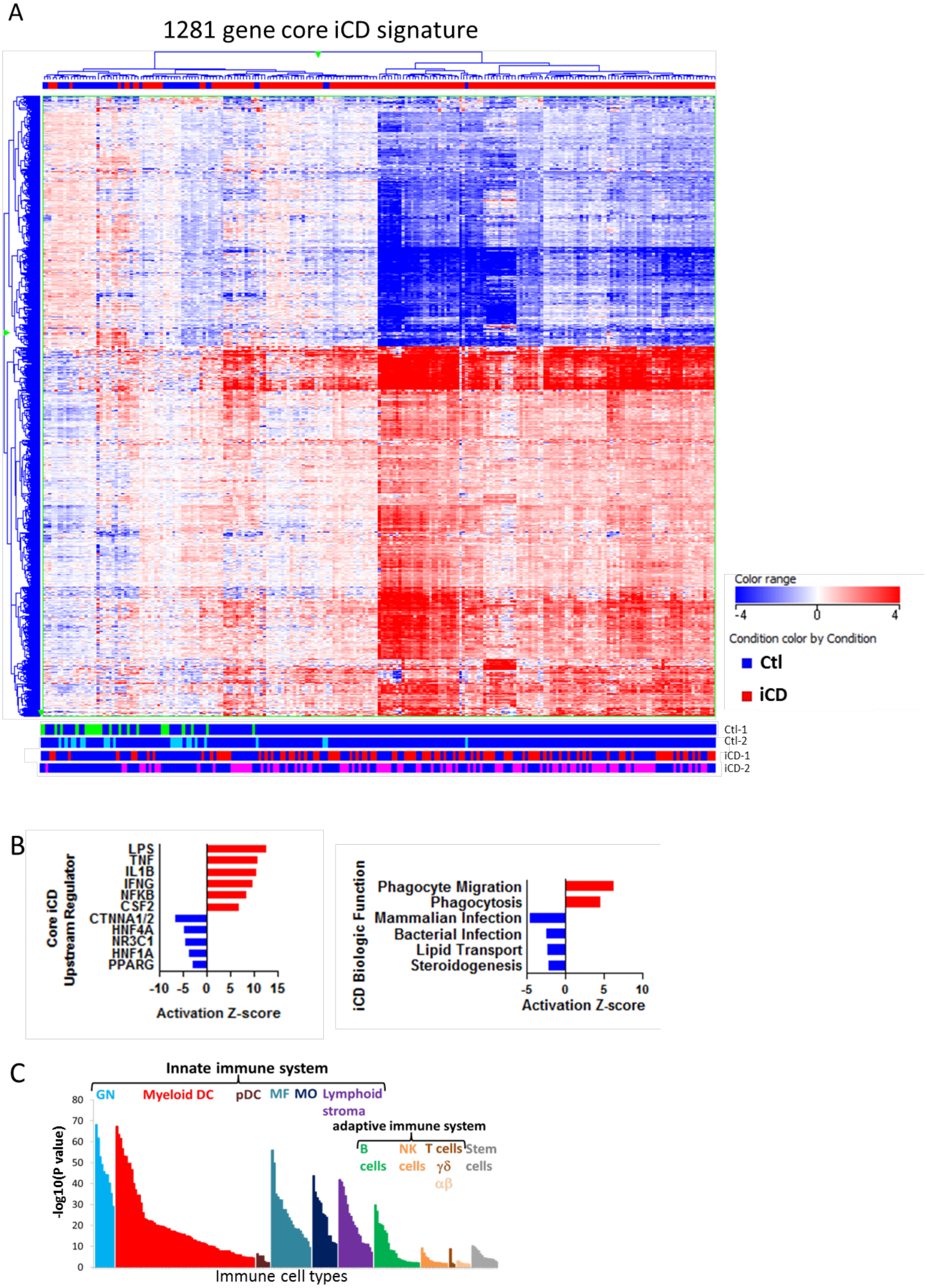
Supplemental Figures



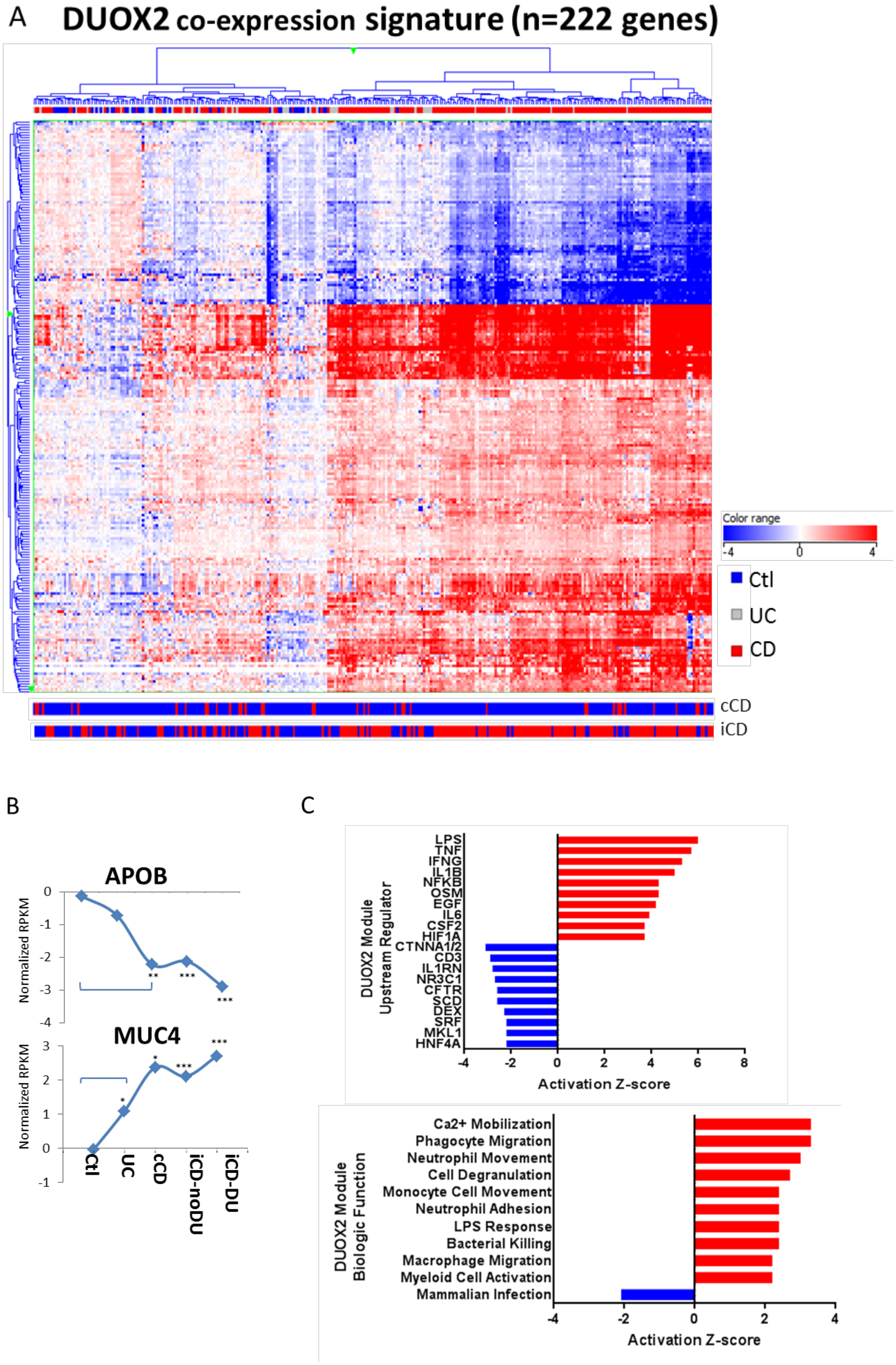
	n
CD	1098
Ctl	380
UC	178
Total	1656

	n
CD	243
Ctl	43
UC	73
Total	359

Supplemental Figure 1. RISK cohort composition and recruitment sites. (**A**) Pie charts of the subject distribution for the overall RISK cohort and the subgroup utilized for the RNA-seq analyses. (**B**) RISK recruitment sites across the United States and Canada.

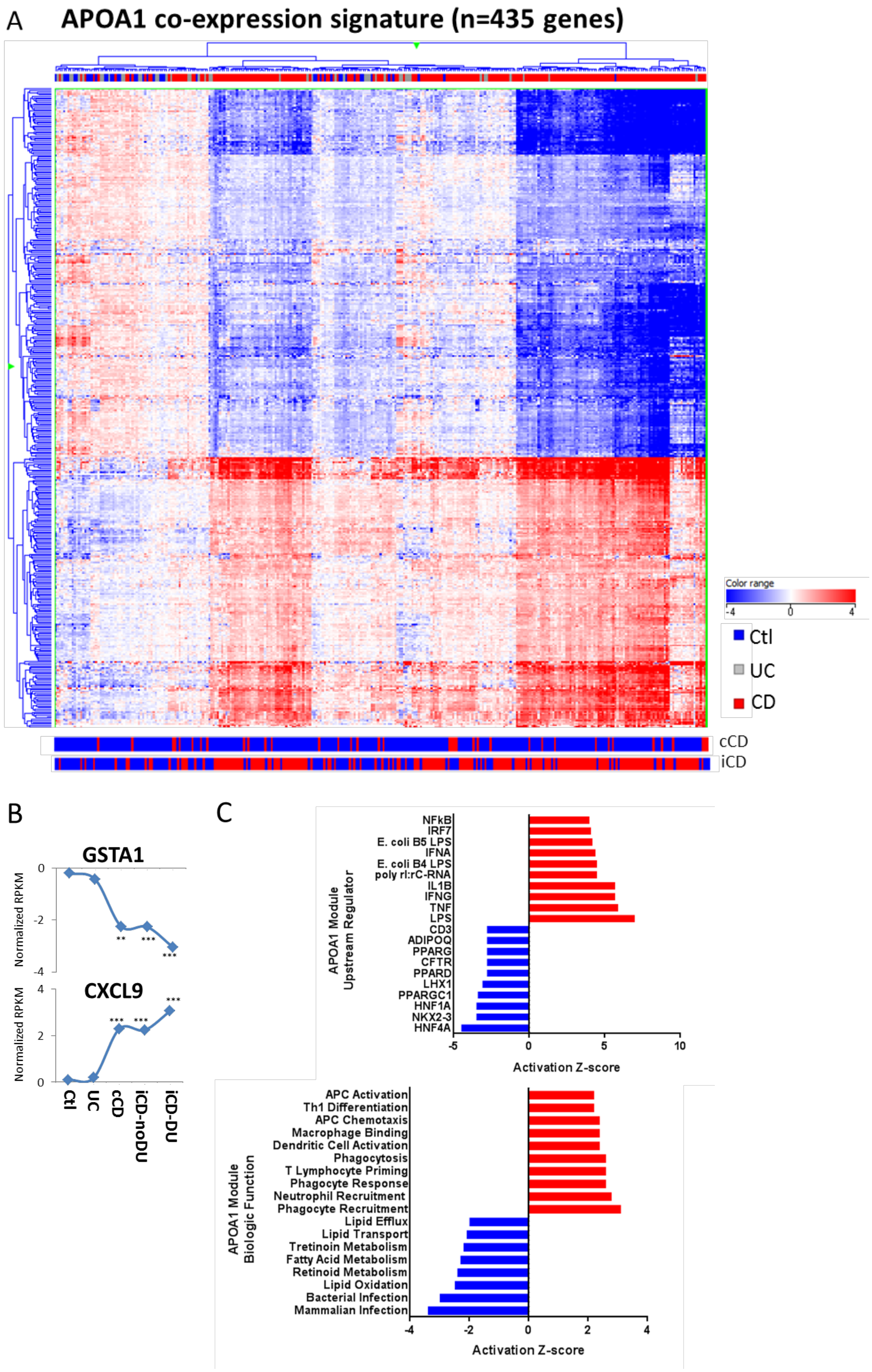

A
1281 gene core iCD signature
Color range
-4
0
4
Condition color by Condition
Ctl
iCD
Ctl-1
Ctl-2
iCD-1
iCD-2
B
Core iCD Upstream Regulator
LPS
TNF
IL1B
IFNG
NFKB
CSF2
CTNNA1/2
HNF4A
NR3C1
HNF1A
PPARG
-10
-5
0
5
10
15
Activation Z-score
iCD Biologic Function
Phagocyte Migration
Phagocytosis
Mammalian Infection
Bacterial Infection
Lipid Transport
Steroidogenesis
-5
0
5
10
Activation Z-score
C
Innate immune system
GN
Myeloid DC
pDC
MF
MO
Lymphoid stroma
adaptive immune system
B cells
NK cells
T cells γδ αβ
Stem cells
80
70
60
50
40
30
20
10
0
-log10(P value)
Immune cell types

Supplemental Figure 2. Core iCD gene signature. (**A**) Hierarchical clustering of the 1,281 genes within the core iCD signature was performed and visualized as a heat map with genes up-regulated compared to control in red and genes down-regulated compared to controls in blue. Above the heat map, individual Ctl (control, blue) and iCD (red) samples are indicated. Below the heat map, Ctl and iCD are further divided into the two independent comparisons that led to the 1,281 as described in Figure 1A. (**B**) Activation *z* scores for upstream regulators (*P* value range: 1E-18 to 4E-89) and biologic functions (*P* value range: 8E-11 to 3E-31) enriched within the core iCD gene signature were determined using Ingenuity Pathways Analysis software (Ingenuity Systems) functional annotation enrichment analyses. (**C**) Immune cell type enrichment of the 762 up-regulated genes within the core iCD gene signature was determined using the Immunological Genome Project data series through ToppGene functional annotation enrichment analyses (1). Ileal enrichment for a given immune cell class (e.g., 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. GN: granulocyte, DC: dendritic cell, MF: macrophage, MO: monocyte.

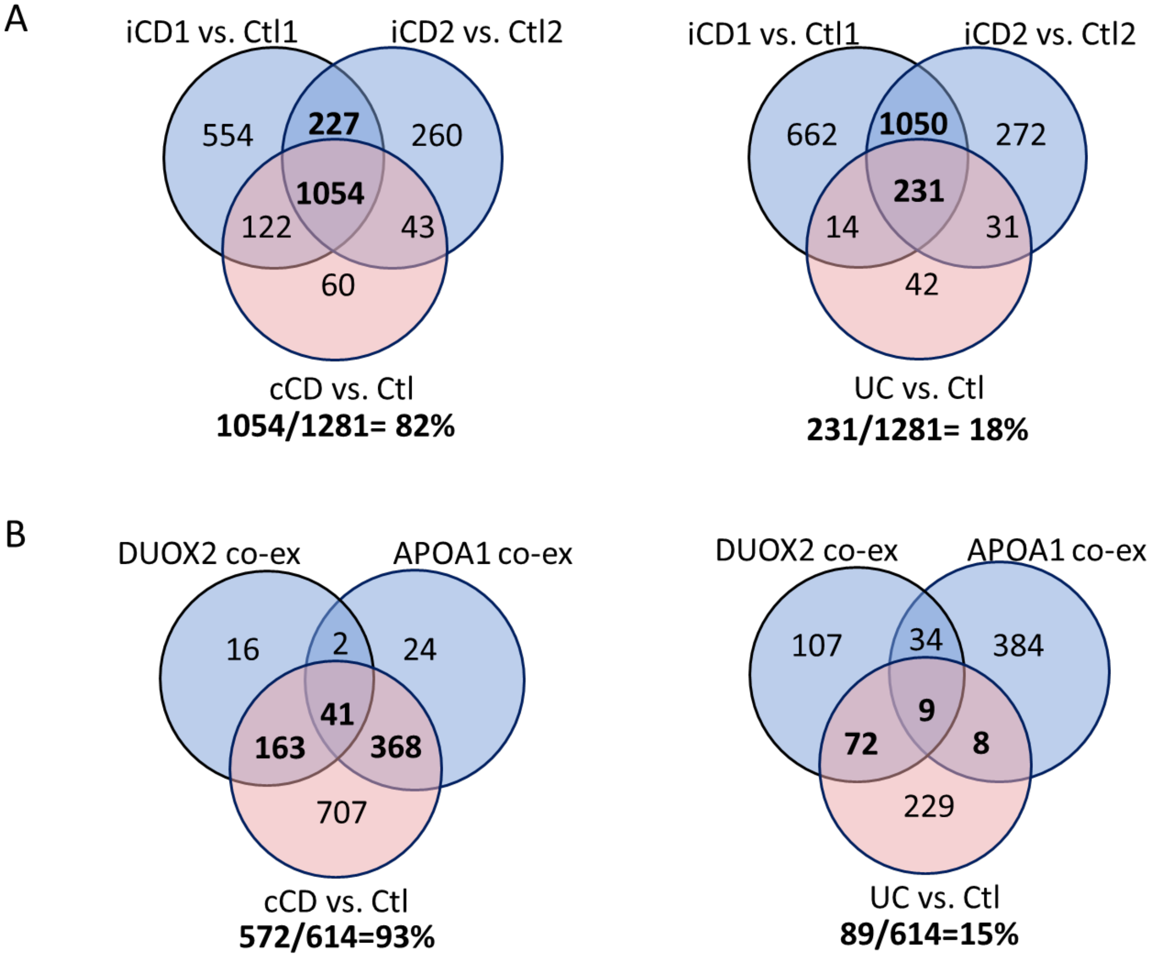


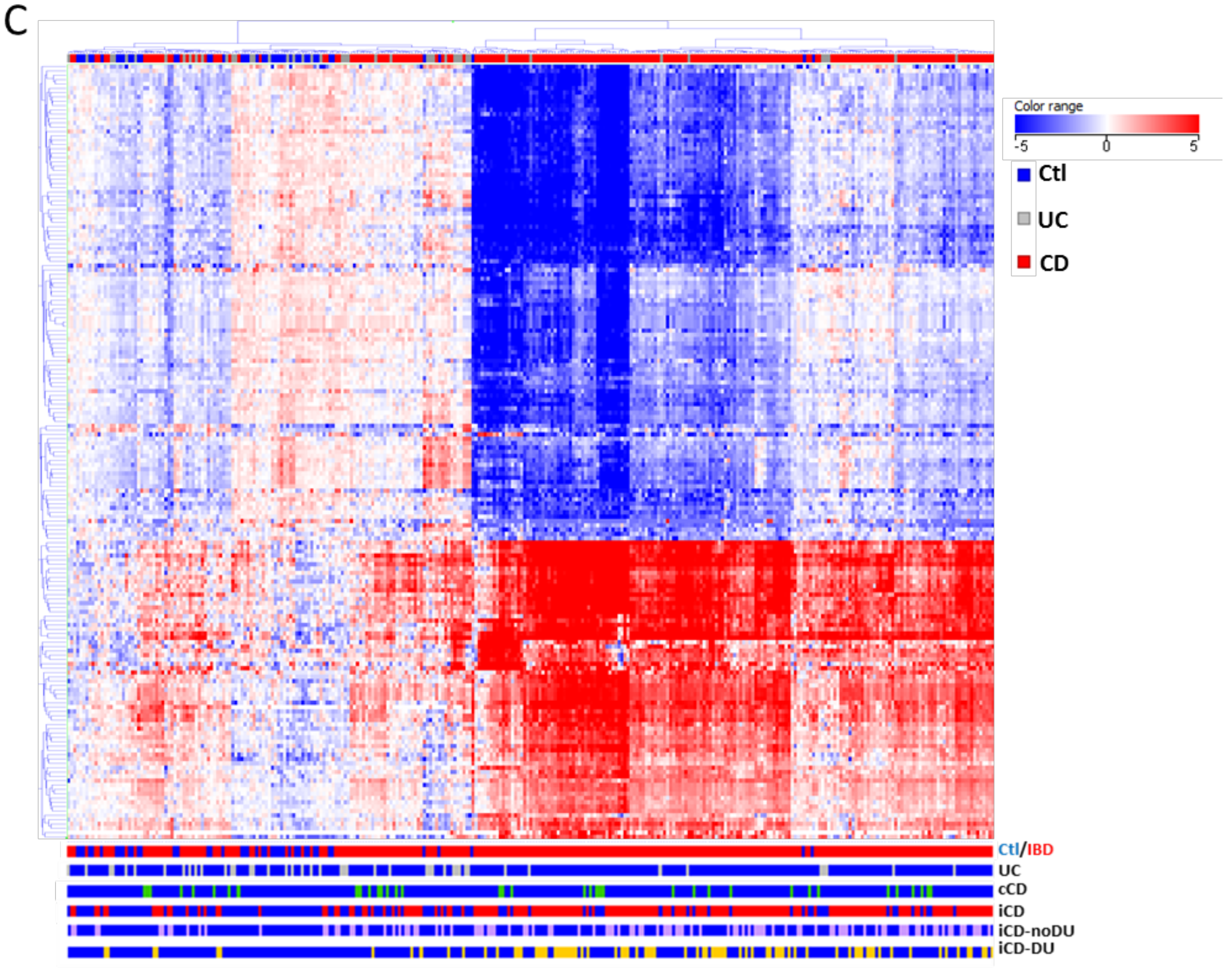
A
DUOX2 co-expression signature (n=222 genes)
Color range
-4
0
4
Ctl
UC
CD
cCD
iCD
B
APOB
Normalized RPKM
0
-1
-2
-3
-4
**
***
***
MUC4
Normalized RPKM
3
2
1
0
-1
*
*
***
***
Ctl
UC
cCD
iCD-noDU
iCD-DU
C
DUOX2 Module Upstream Regulator
LPS
TNF
IFNG
IL1B
NFKB
OSM
EGF
IL6
CSF2
HIF1A
CTNNA1/2
CD3
IL1RN
NR3C1
CFTR
SCD
DEX
SRF
MKL1
HNF4A
-4
-2
0
2
4
6
8
Activation Z-score
DUOX2 Module Biologic Function
Ca2+ Mobilization
Phagocyte Migration
Neutrophil Movement
Cell Degranulation
Monocyte Cell Movement
Neutrophil Adhesion
LPS Response
Bacterial Killing
Macrophage Migration
Myeloid Cell Activation
Mammalian Infection
-4
-2
0
2
4
Activation Z-score

Supplemental Figure 3. Heat map and upstream regulators for the *DUOX2* gene co-expression signature. (**A**) Hierarchical clustering of the 222 genes contained in the *DUOX2* gene co-expression signature was performed and visualized as a heat map with genes up-regulated compared to control in red and genes down-regulated compared to control in blue. Individual Ctl (blue), UC (grey), and iCD (red) sample results are indicated above the heat map. Below the heat map, individual cCD and iCD within the overall CD groups are indicated in red. (**B**) Average *APOB* and *MUC4* gene expression across the clinical subgroups is shown. Differences between patient subgroups were tested using Kruskal-Wallis with Dunn's Multiple Comparison test of all groups vs. Ctl. **P* < 0.05, ***P* < 0.01, ****P* < 0.001. (**C**) Activation *z* score (*P* value range: 3E-3 to 5E-18) of upstream regulator and biologic function enrichment of the *DUOX2* module genes was determined using IPA functional annotation enrichment analyses (Ingenuity Systems).


A
APOA1 co-expression signature (n=435 genes)
Color range
-4
0
4
Ctl
UC
CD
cCD
iCD
B
GSTA1
Normalized RPKM
0
-2
-4
**
***
***
CXCL9
Normalized RPKM
4
2
0
***
***
***
Ctl
UC
cCD
iCD-noDU
iCD-DU
C
APOA1 Module
Upstream Regulator
NFkB
IRF7
E. coli B5 LPS
IFNA
E. coli B4 LPS
poly rI:rC-RNA
IL1B
IFNG
TNF
LPS
CD3
ADIPOQ
PPARG
CFTR
PPARD
LHX1
PPARGC1
HNF1A
NKX2-3
HNF4A
-5
0
5
10
Activation Z-score
APOA1 Module
Biologic Function
APC Activation
Th1 Differentiation
APC Chemotaxis
Macrophage Binding
Dendritic Cell Activation
Phagocytosis
T Lymphocyte Priming
Phagocyte Response
Neutrophil Recruitment
Phagocyte Recruitment
Lipid Efflux
Lipid Transport
Tretinoin Metabolism
Fatty Acid Metabolism
Retinoid Metabolism
Lipid Oxidation
Bacterial Infection
Mammalian Infection
-4
-2
0
2
4
Activation Z-score

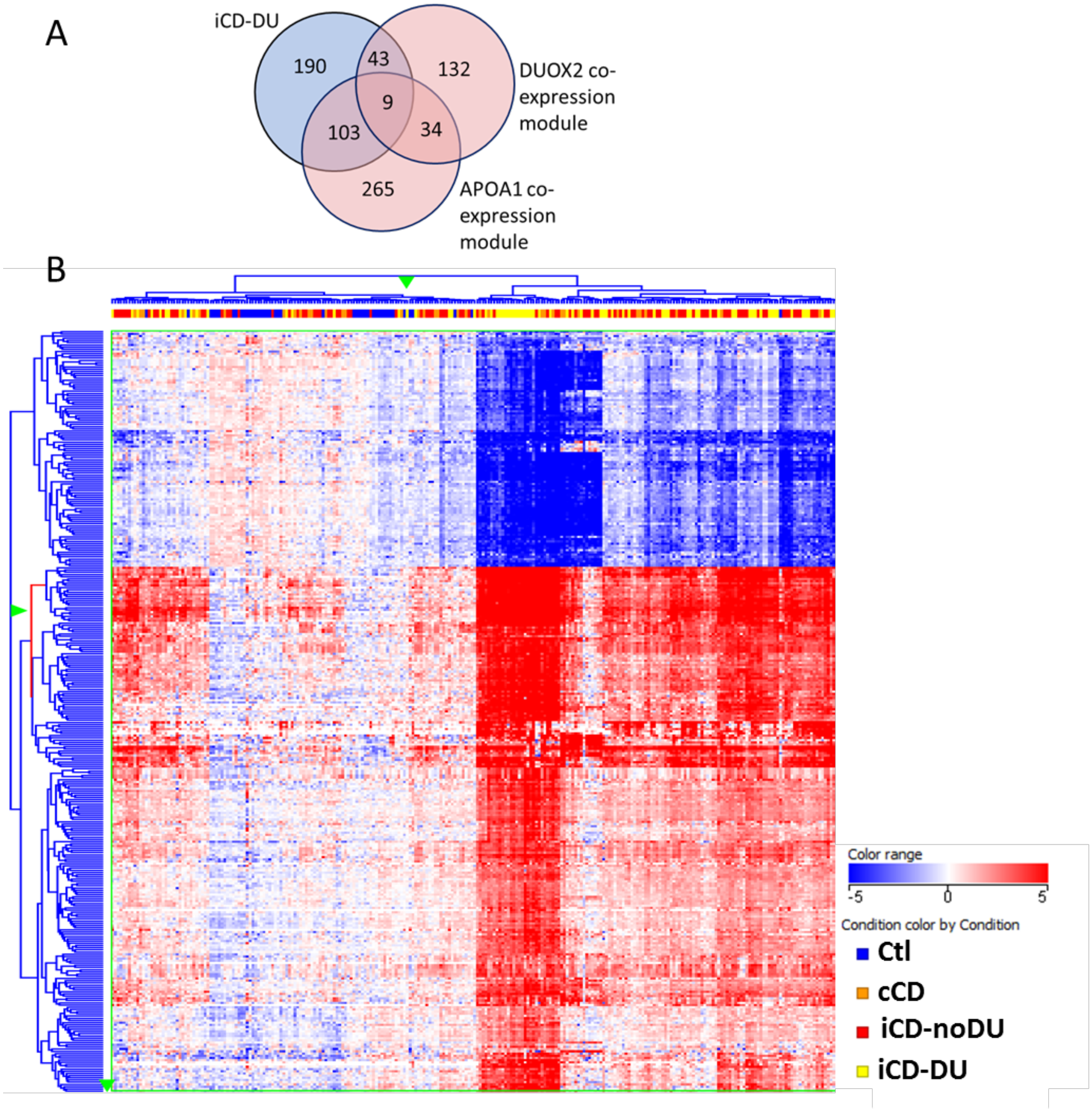
Supplemental Figure 4. Heat map and upstream regulators for the *APOA1* gene co-expression signature . (**A**) Hierarchical clustering of the 435 genes contained in the *APOA1* gene co-expression signature was performed and visualized as a heat map with genes up-regulated compared to control in red and genes down-regulated compared to control in blue. Individual Ctl (blue), UC (grey), and iCD (red) sample results are indicated above the heat map. Below the heat map, individual cCD and iCD within the overall CD groups are indicated in red. (**B**) Average *GSTA1* and *CXCL9* gene expression across the clinical subgroups is shown. Differences between patient subgroups were tested using Kruskal-Wallis with Dunn's Multiple Comparison test of all groups vs. Ctl. **P* < 0.05, ***P* < 0.01, ****P* < 0.001. (**C**) Activation *z* score (*P* value range: 1E-6 to 3E-28) of upstream regulator and biologic function enrichment of the *APOA1* module genes as determined using IPA functional annotation enrichment analyses (Ingenuity Systems).


A
iCD1 vs. Ctl1
iCD2 vs. Ctl2
554
227
260
1054
122
43
60
cCD vs. Ctl
1054/1281= 82%
iCD1 vs. Ctl1
iCD2 vs. Ctl2
662
1050
272
231
14
31
42
UC vs. Ctl
231/1281= 18%
B
DUOX2 co-ex
APOA1 co-ex
16
2
24
41
163
368
707
cCD vs. Ctl
572/614=93%
DUOX2 co-ex
APOA1 co-ex
107
34
384
9
72
8
229
UC vs. Ctl
89/614=15%

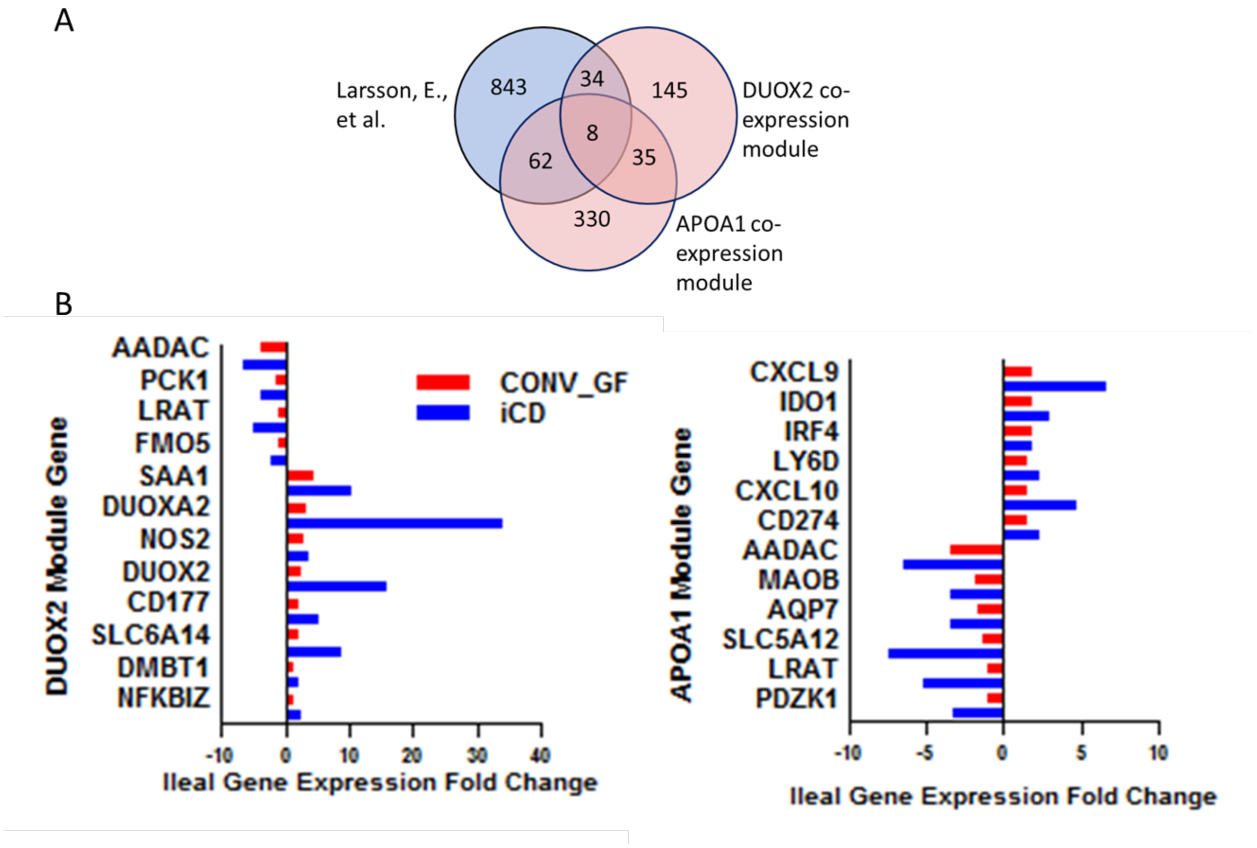


Supplemental Figure 5. *APOA1* and *DUOX2* co-expression signature genes are enriched in the genes that distinguish cCD from UC. (**A**) Left: Venn diagram shows the overlap of 1,054 genes differentially expressed between cCD and Ctl (fold change of 1.5) and genes within the core 1,281 core iCD gene list (83%). Right: Venn diagram shows the overlap of 231 genes differentially expressed between UC and Ctl (fold change of 1.5) and genes within the core 1,281 core iCD gene list (18%). (**B**) Left: Venn diagram shows the overlap of 572 genes differentially expressed between cCD and Ctl (fold change of 1.5) and genes within the *DUOX2* and *APOA1* gene co-expression signatures. The overlapped 572 genes are 93% of the total 614 genes comprising the combined *APOA1* and *DUOX2* gene co-expression signatures. Right: Venn diagram shows the overlap of 89 genes differentially expressed between UC and Ctl (fold change of 1.5) and genes within the *DUOX2* and *APOA1* gene co-expression signatures. The overlapped 89 genes are 15% of the total 614 genes comprising the combined *APOA1* and *DUOX2* gene co-expression signatures. (**C**) Unsupervised hierarchical clustering of the 179 genes common to the cCD versus UC (fold change

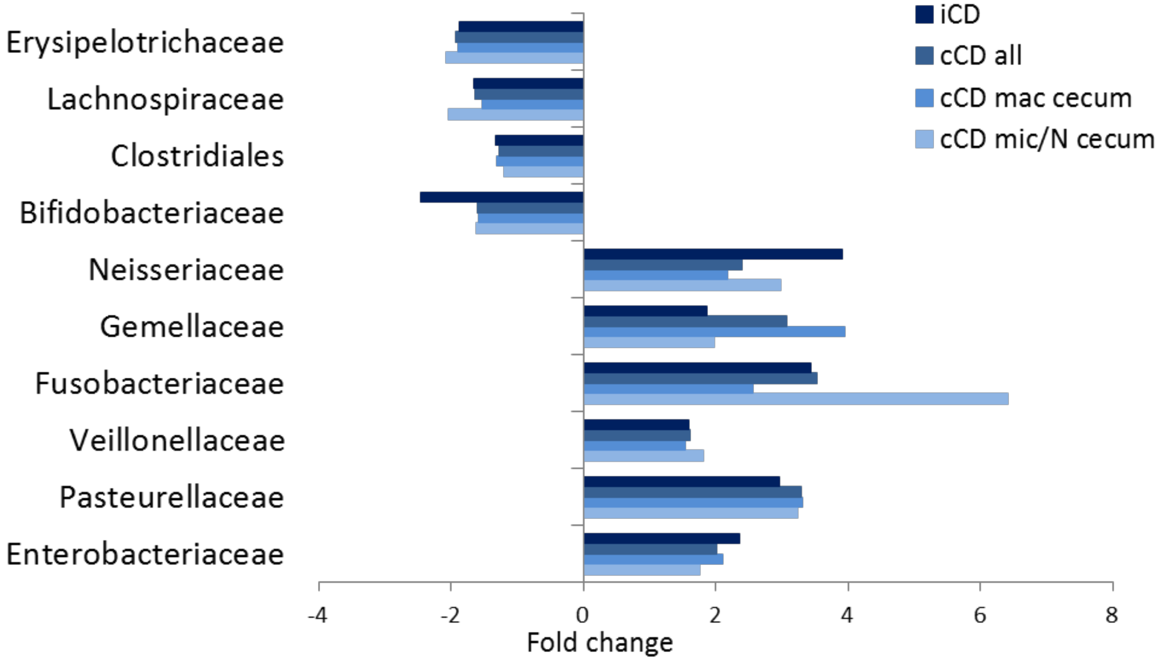
≥ 2) and the *APOA1/DUOX2* gene co-expression signatures is shown in the heat map. Genes up-regulated compared to control are in red and genes down-regulated compared to controls are in blue. Individual Ctl (blue) UC (grey), and iCD (red) samples are indicated above the heat map. Below the heat map, individual samples from the different groups (Ctl, all IBD, UC training set, cCD training set, iCD-noDU, and iCD-DU) are shown.



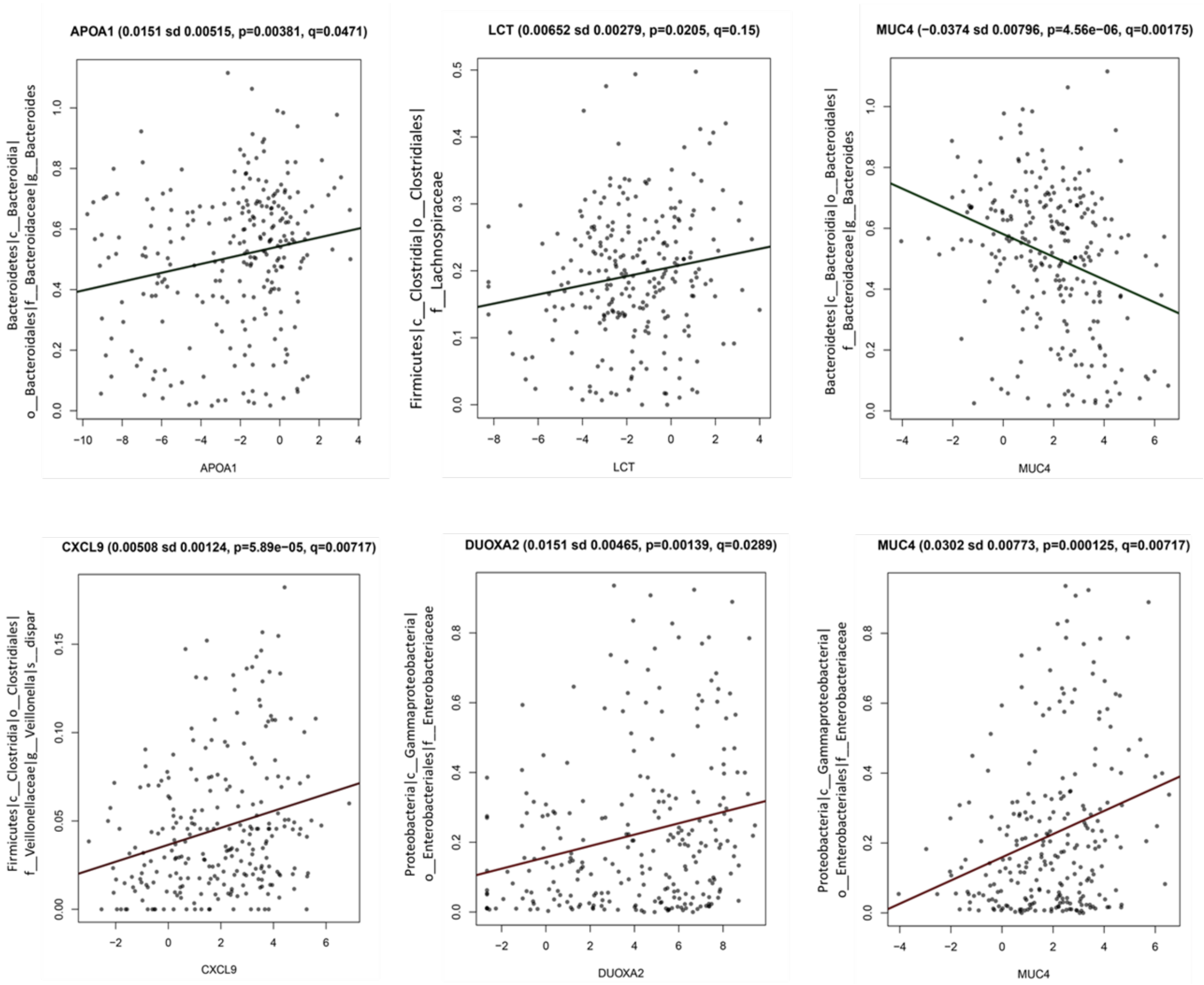
Supplemental Figure 6. Heat map for the iCD-DU gene list. (**A**) The Venn diagram shows the overlap between the iCD-DU gene list and the *DUOX2* and *APOA1* gene co-expression signatures. (**B**) A heat map for hierarchical clustering of the 345 genes contained in the iCD-DU gene list with genes up-regulated compared to control in red and genes down-regulated compared to control in blue is shown. Individual Ctl (blue), cCD (orange), iCD-noDU (red), and iCD-DU (yellow) sample results are indicated.



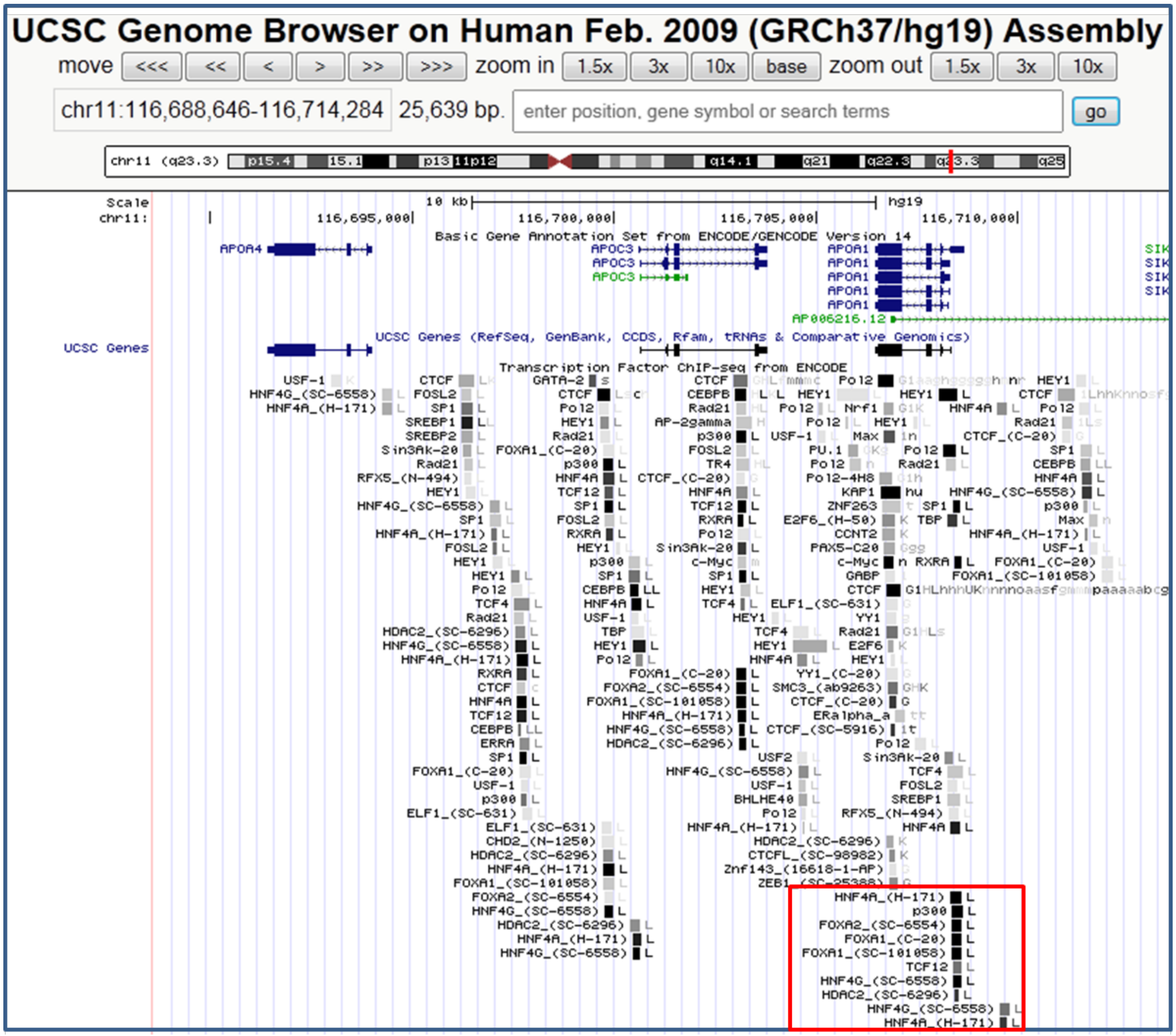
Supplemental Figure 7. *DUOX2* and *APOA1* co-expression signature genes are regulated by bacterial colonization in mice. (**A**) The Venn diagram shows the overlap between ileal genes whose expression changed after bacterial colonization in mice (2) and genes within the *DUOX2* and *APOA1* gene co-expression signatures. (**B**) Bar graphs show changes in ileal gene expression in mice following bacterial colonization (conventional/colonized compared to germ-free, CONV_GF) and iCD for selected genes from the *DUOX2* and *APOA1* gene co-expression signatures.



Supplemental Figure 8. The ileal microbial community in cCD patients with inflamed or non-inflamed cecum. Fold change for each taxa was calculated by dividing the mean abundance in the cases [cCD (54 patients) or iCD (226 patients)] by that of the controls (154 patients) and is shown for microbiota with differential abundance in CD compared to Ctl. The cCD group was further subdivided to cCD with macroscopically inflamed cecum (cCD mac cecum, 37 patients) or those with normal appearing cecum but with either abnormal histological feature or normal histology (cCD mic/N cecum. 15 patients).



Supplemental Figure 9. Reduction of Firmicutes species and expansion of selected Proteobacteria are associated with changes in expression of genes from the *APOA1* and *DUOX2* gene co-expression signatures. The association between specific microbial taxa abundance and ileal gene expression as determined using MaAsLin is shown. The r (sd) coefficient effect size, and *P* and *q* value tests for significance, are shown above each scatter gram.



UCSC Genome Browser on Human Feb. 2009 (GRCh37/hg19) Assembly

chr11:116,688,646-116,714,284 25,639 bp.

Supplemental Figure 10. ENCODE transcription factor CHIP-seq output. The rectangle shows HNF4A, HNF4G, as well as other transcription factors binding within the *APOA1/APOC3/APOA4* locus.

Supplemental References

1. Chen, J., Bardes, E.E., Aronow, B.J., and Jegga, A.G. 2009. ToppGene Suite for gene list enrichment analysis and candidate gene prioritization. *Nucleic Acids Res* 37:W305-311.
2. Larsson, E., Tremaroli, V., Lee, Y.S., Koren, O., Nookaew, I., Fricker, A., Nielsen, J., Ley, R.E., and Backhed, F. 2012. Analysis of gut microbial regulation of host gene expression along the length of the gut and regulation of gut microbial ecology through MyD88. *Gut* 61:1124-1131.

Table 1. RISK RNA-seq Cohort Clinical and Demographic Characteristics.

	Ctl n=43	**UC 1 n=41**	**cCD 1 n=34**	**UC 2 n=28**	**cCD 2 n=24**	**iCD 1 n=89**	**iCD 2 n=88**	**all iCD n=177**	**iCD DU n=77**	**iCD noDU n=100**
Mean (SD) Age (years)	11(3)	12(3)	12(3)	13(4)	13(3)	12(3)	12(3)	12(3)	12(3)	12(3)
Male gender (%)	65	46	53	71	50	60	63	62	57	65
MED ethnicity (3 of 4 grandparents) (%)	97	85	91	86	83	90	88	89	91	88
Perianal involvement (%)	0	0	18	0	27	19	17	18	19	17
Ileal deep ulcers (%)	0	0	0	0	0	45	42	43	100	0
Body mass index Z<-2 (%)	3	2	18	4	13	28	16	22	23	21
PCDAI at diagnosis										
≤10 (inactive, %)	na	na	9	na	9	8	10	9	12	7
11 to 30 (mild, %)	na	na	39	na	36	36	46	41	32	47*
>30 (moderate-severe, %)	na	na	52	na	55	56	44	50	56	46

Differences between selected groups were tested by ANOVA for continuous variables and Chi-square for dichotomous variables. MED: mixed European descent, PCDAI: Pediatric Crohn Disease Activity Index. **P* =0.045 vs iCD DU.