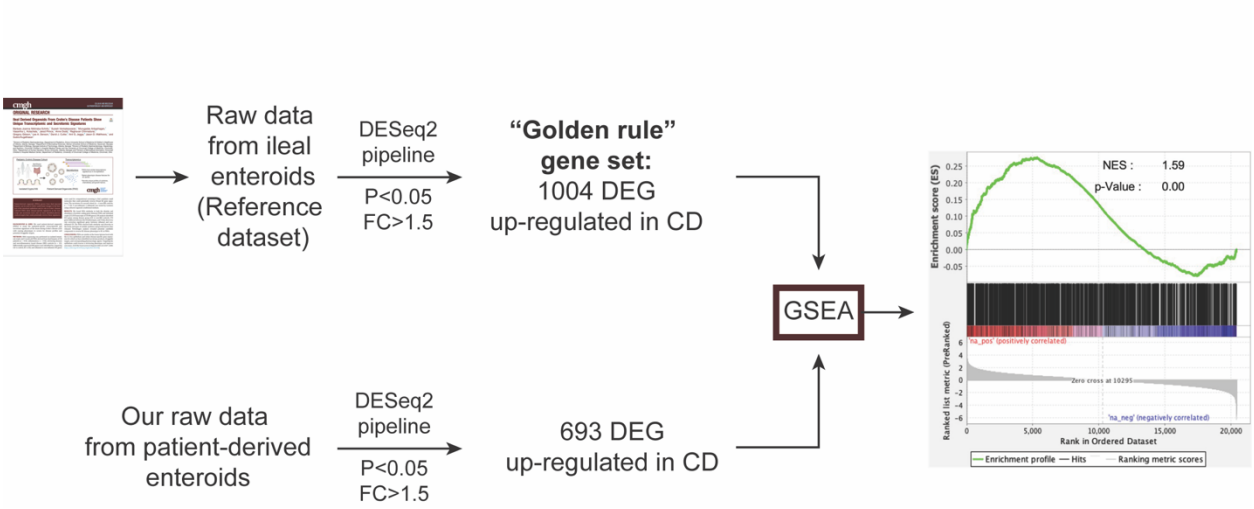
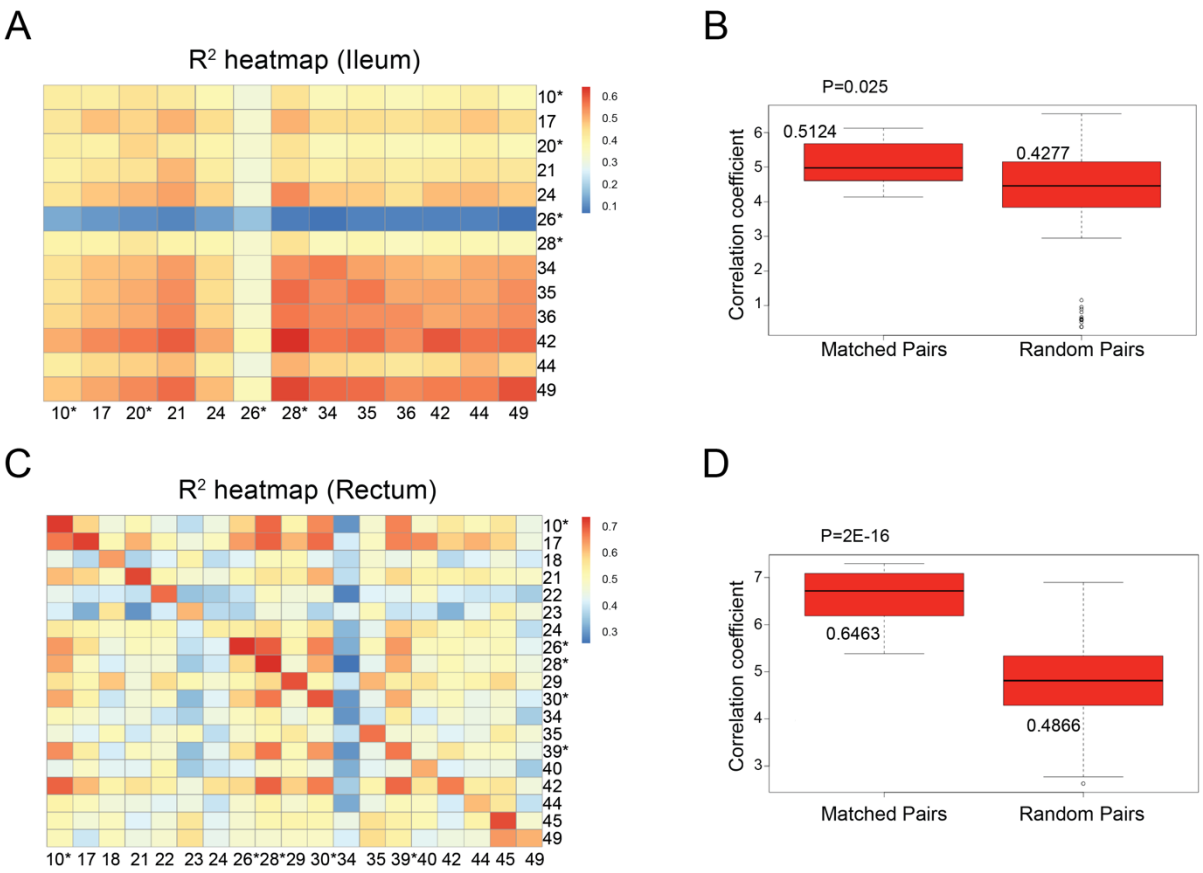


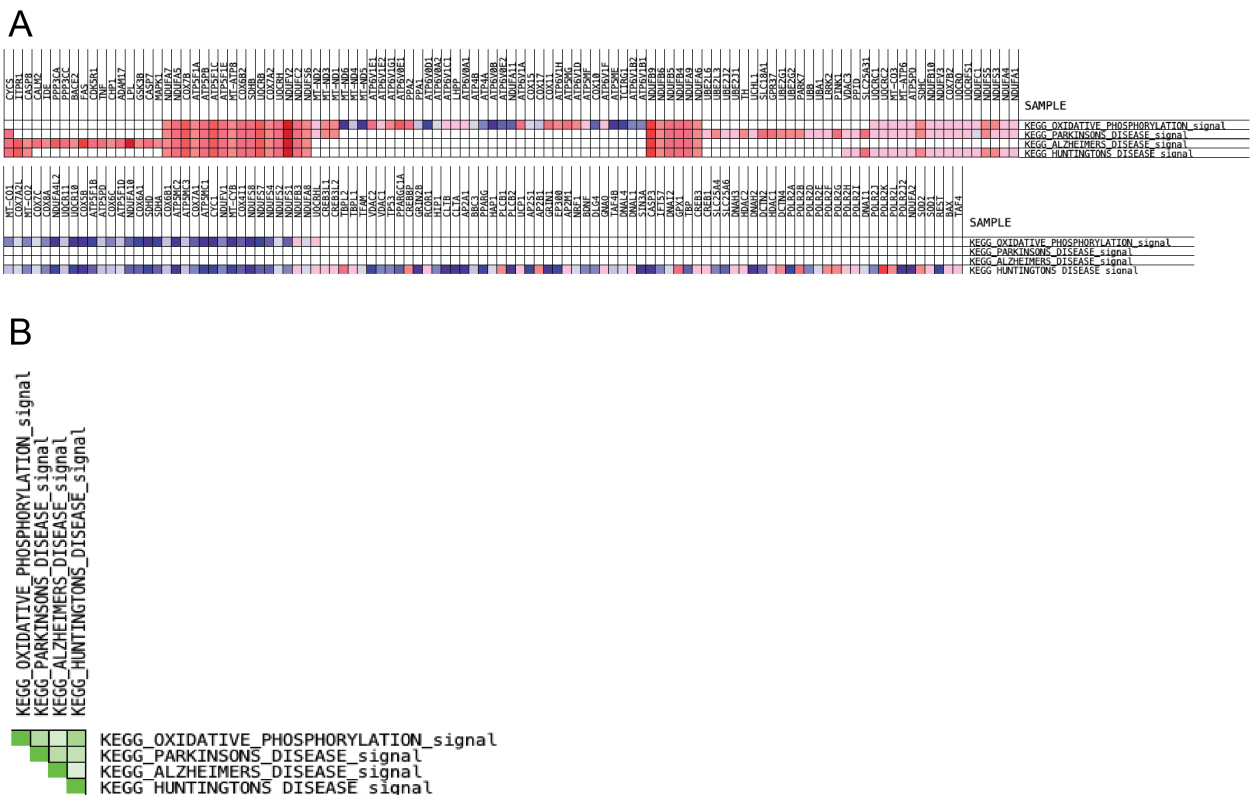
Supplementary figure 1. Pipeline and data validation against a reference gene set

(**A**) Pipeline validation schematic: the raw data accompanying the paper by Niclinska-Schirtz et al were download from EMBL-EBI ArrayExpress database and then processed with our own DESeq2 pipeline, yielding the "Golden rule" gene set comprising 1004 genes differentially expressed in Crohn's disease enteroids. Gene set enrichment analysis of our list of 693 genes differentially expressed in enteroids from Crohn's disease is performed against the reference dataset, and the plot (**B**) shows strong agreement between the two datasets.



Supplementary figure 2. Transcriptomic features are retained in enteroid/colonoid-biopsy pairs matched by subject

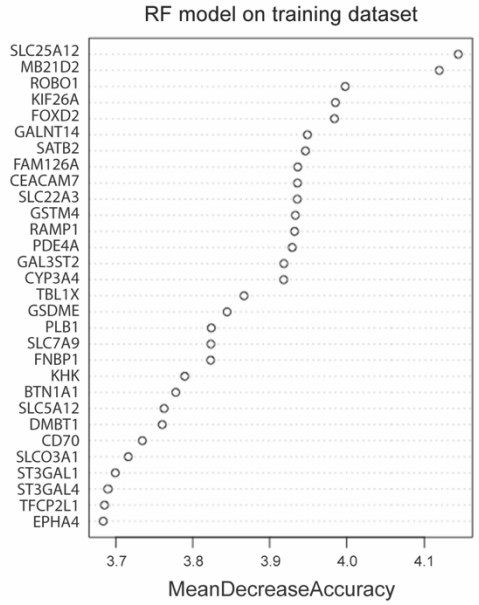
Heatmap plots of correlation coefficients comparing the normalized gene expression matrices for all the genes from individual biopsies (ileal **(A)** or rectal **(C)**) to enteroids (**A**) or colonoids (**C**). Asterisks indicate Crohn's disease subjects. Average correlation coefficients and P-values for subject-matched vs random pairs are presented in panels **(B)** (Ileum) and **(D)** (rectum).



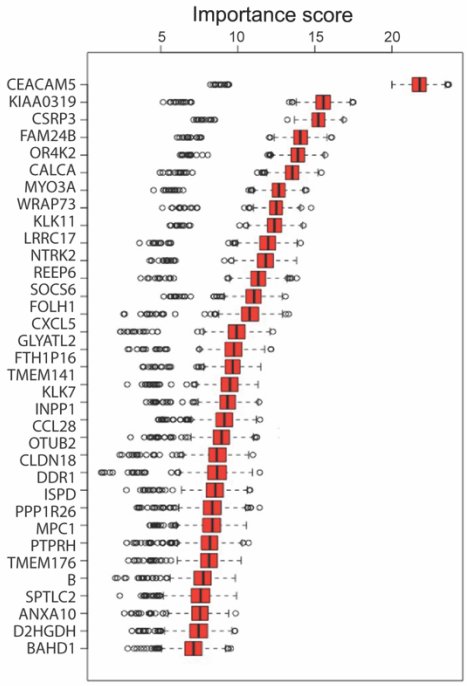
Supplementary figure 3. Neurodegenerative disease KEGG pathways share genes with the Oxidative Phosphorylation pathway

(A) Visual representation of the leading-edge analysis for the Parkinson's disease, Alzheimer's disease, and Huntington's disease pathways, as well as the Oxidative phosphorylation pathway. There is a pronounced overlap in genes that account for enrichment of the gene sets in our data. (**B**) Graphical representation of the overlap between the gene sets for the 4 pathways. The color intensity indicates the magnitude of overlap between the two subsets. A dark green cell indicates that the two gene sets have the same leading-edge genes, while a white cell would indicate no leading edge genes in common.

A



B



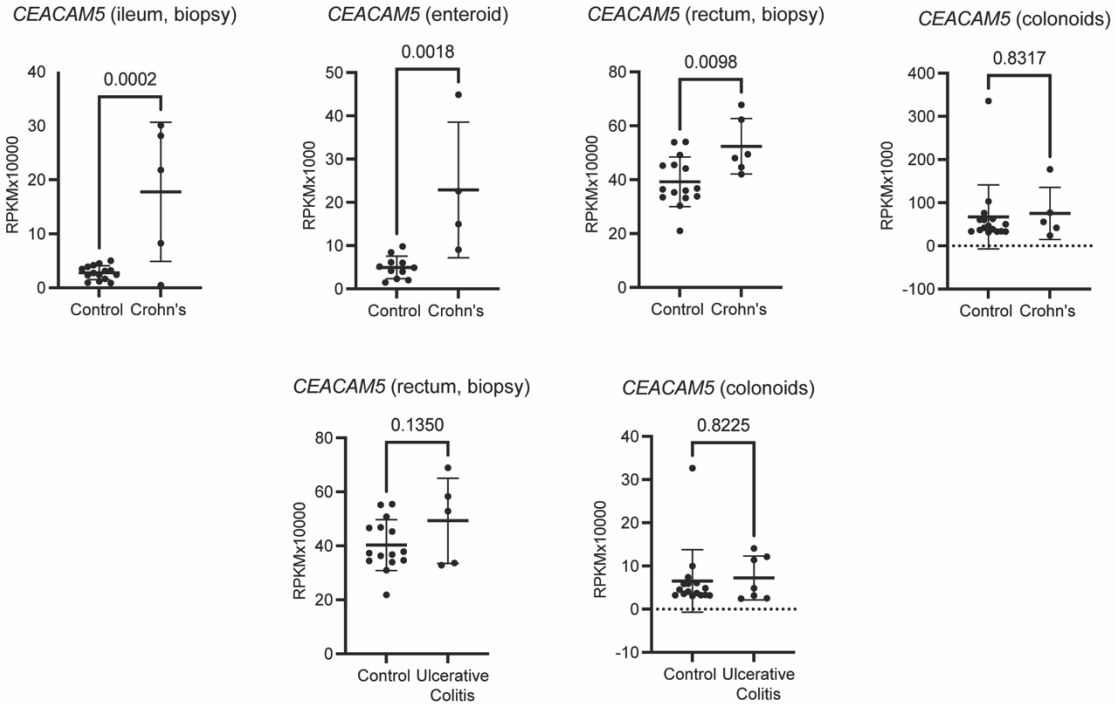
Statistical summary

Accuracy	1
Balanced Accuracy	1
Kappa	1
P-Value	9.31e-10

SVM model on training dataset

Accuracy	0.9250
Kappa	0.846
Accuracy SD	0.138
Kappa SD	0.281

C

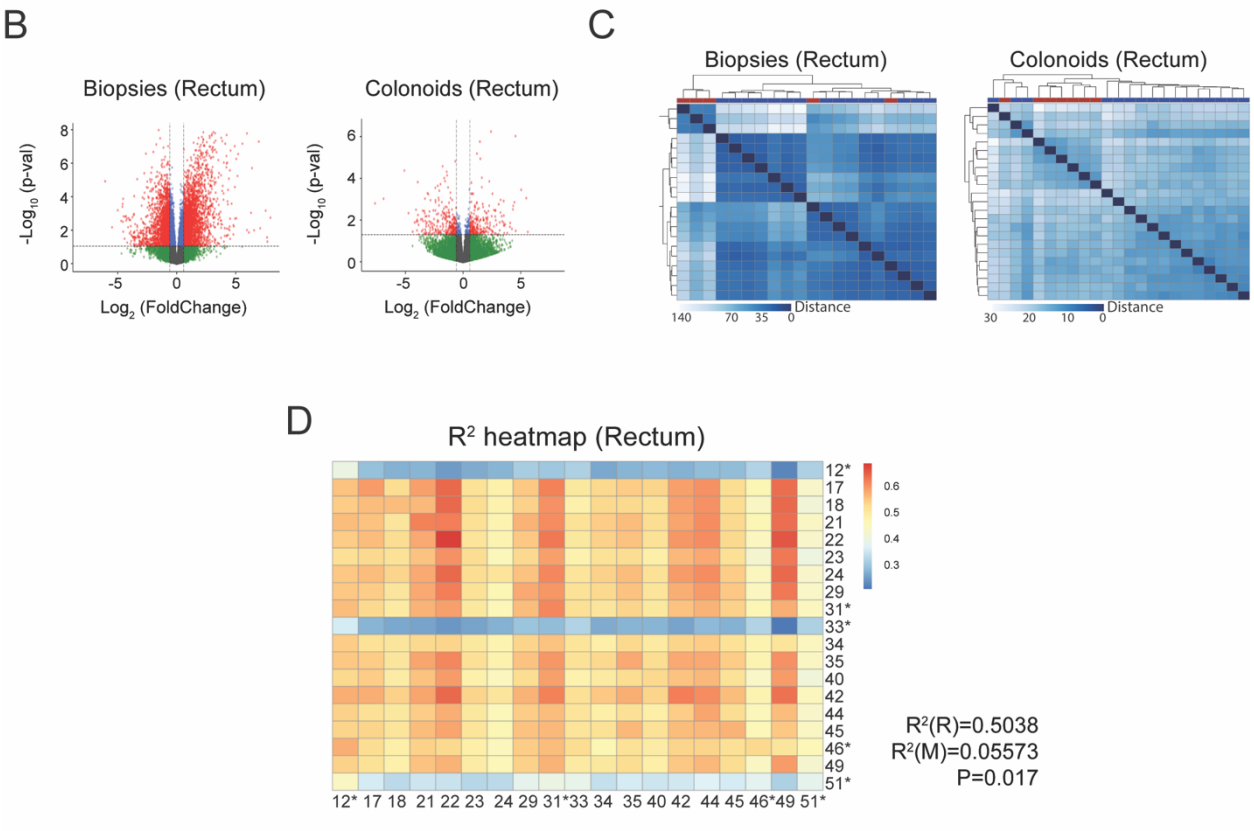


Supplementary figure 4. Characteristics of the random forest model classifiers

(A) Discriminating features of the “tissue of origin classifier”: 30 genes most important based on the mean decrease accuracy score. (**B**) Discriminating features of the “disease state classifier”: 30 genes most important based on the importance score. **(C)** Reads per kilobase per million (RPKM) values for CEACAM5 gene from RNA-seq of corresponding specimens.

A

Ulcerative colitis patients	
N	6
Male, n (%)	3 (43%)
Age at sample collection, years, mean (±SD)	15.1 (2.1)
Age range	11-17
Pre-existing IBD diagnosis, n (%)	1 (17%)
Treatment at the time of biopsy, n (%)	1 (17%)



Supplementary figure 5. Rectal colonoids retain transcriptomic differences of rectal biopsies from ulcerative colitis patients

(A) Ulcerative colitis patient demographics. The one patient who had a pre-existing IBD diagnosis was receiving treatment (infliximab and methotrexate) at the time of colonoscopy. **(B)** Volcano plots demonstrating the changes in gene expression between the control and Crohn's disease samples from rectal biopsies and colonoids. Red and green dots represent the genes that are significantly upregulated or downregulated ($|FC| \geq 1.5$ and p value ≤ 0.05) in ulcerative colitis samples compared to controls, respectively. Grey dots represent the genes with no significant expression differences **(C)** Hierarchical clustering of top 200 differentially expressed genes selected (ulcerative colitis versus control), generated using the 100 differentially expressed genes from the upregulated gene list and 100 from downregulated gene list. **(D)** Heatmap plot of correlation coefficients comparing the normalized gene expression matrices for all the genes from individual biopsies to colonoids. Asterisks indicate ulcerative colitis subjects.