Supplementary Methods:

Ethical approval

Ethical approval was obtained from the local research committee (REC 12/EE/0482 and REC 17/EE/0265) and patients were prospectively recruited following informed patient and/or carer consent. All investigations were carried out according to the Declaration of Helsinki and Good Clinical Practice Guidelines.

Patient and Public Involvement

This study is part of an ongoing translational research theme (TRIPP study – **T**ranslational **R**esearch in **I**ntestinal **P**hysiology and **P**athology). As part of this theme, we have established a parent and patient support group who is actively involved in the development of our clinical service as well as research strategies. Regular interactions between the clinical team, members of the laboratory group with patients and parents include a bi-monthly research newsletter as well as family events.

Patient recruitment and clinical data recording

All patients recruited to this study were followed for a minimum of 18 months post-diagnosis in the Cambridge pediatric gastroenterology unit, and detailed clinical data was prospectively recorded. This included phenotypic parameters at diagnosis (e.g. presence of perianal disease) and information on disease course and outcomes. The latter covered number of treatment escalations, treatment history for surgical intervention, treatment with biologics, and use of immunomodulator (i.e. Azathioprine). To account for the fact that disease outcome in patients suffering from IBD is not restricted to a single measure, we also calculated a compound severity score. The score, which has been previously published (with minor modifications), considered number of treatment escalations, escalation to treatment with biologics, presence of peri-anal disease, and the requirement of IBD related surgery[1].

Briefly, the score was calculated at 12 months from diagnosis and patients were categorised into mild (full response to single induction treatment, maintenance treatment with 5AZA for UC and azathioprine for CD, no further escalation), severe (requirement for escalation to second line treatment with biologics, requirement for IBD related surgery to control symptoms, more than 2 additional induction treatments post diagnosis), moderate (remaining cases). All patients were treated according to a standard protocol and treatment decisions were made by a

multi-disciplinary team resulting in highly comparable treatment courses and outcome measures.

Sample Collection

At diagnostic endoscopy, mucosal biopsies were obtained from proximal and distal small bowel (i.e. Duodenum = DUO, Terminal Ileum = TI) and distal large bowel (i.e. Sigmoid Colon = SC). Inflammation status of a mucosal sample (inflamed vs. non-inflamed) was based on histological assessment of paired samples taken within 2 cm. Histological assessment was performed by an experienced histopathologist. Biopsies were processed immediately for the generation of intestinal epithelial organoids or molecular profiling.

Human intestinal epithelial organoid culture (IEOs) generation and biobank

Intestinal crypts were isolated from human mucosal biopsies and cultured in a growth medium, as previously described[1, 2, 3], to generate human intestinal epithelial organoids (IEOs). Briefly, freshly obtained biopsies were washed with PBS and incubated in 2.5mM EDTA at 4oC for 30 minutes on a roller. Biopsies were washed to eliminate residual EDTA and crypts were released by pipetting followed by centrifugation, seeding in Matrigel (Corning) and covered in complete organoid medium (composition provided in Supplementary Table x). Organoids were maintained by medium change every 2 days and split every 7-10 days. Following expansion of IEOs over a minimum of two weeks in culture, frozen stocks were generated and stored in a living biobank and/or further expanded for experiments including cytokine stimulation. IEOs were frozen after centrifugation at 800g for 5 minutes, resuspension of the pellet in 1ml of Recovery Cell Culture Freezing Medium (Gibco), placement in a cryo-freezing container (“Mr. Frosty,” Thermo Fisher Scientific) for at least 24h, and then transfer to -150°C for long-term storage.

Each organoid line is accompanied by detailed clinical information, quality control data including bright field images (taken using an EVOS FL system (Life Technologies)) to illustrate organoid growth, cell numbers as well as culture conditions used. A summary of all samples is provided in Table S1.

Assessment of organoid growth

The growth of IEOs was monitored using the Incucyte SX5 system (Sartorius AG, Göttingen, Germany) by imaging them every 6 hours over 6 days. Following the experiment, image analysis was performed using the Incucyte IEO analysis software to measure IEOs area over

time. Each experimental group, including Controls and Chron's Disease samples, consisted of two lines of IEOs. For each line, images were captured and analysed from three separate wells to derive an average measurement of IEO growth.

RNA/DNA extraction, Reverse transcription and quantitative PCR

DNA and/or RNA from human or mouse organoids were extracted using the AllPrep DNA/RNA mini kit. RNA was reverse-transcribed and used to perform a quantitative PCR (qPCR) as previously reported[4].

***In vitro* stimulation of human and murine IEOs with inflammatory cytokines**

Unless otherwise indicated, experiments with unmodified paediatric IEOs lines were cultured in a 48-well-plate for 5 days after splitting prior to the addition of recombinant human TNFα (H8916; Sigma Aldrich, Burlington, MA) at 40 ng/mL; IFNγ (PHC4031; Life Technologies, Carlsbad, CA) at 20 ng/mL, or no treatment for 24 hours as previously described[5]. Organoids were then harvested for concurrent DNA and RNA extraction.

For experiments involving PB NLRC5-mCherry IEOs, organoids were grown for 5 days after splitting, then *NLRC5* overexpression induced by stimulation with 1 μg/ml doxycycline hyclate or vehicle control for 24 hours. Secondary treatment with IFNγ, TNFα or medium alone was then administered for 24 hours as described above, before harvesting for RNA extraction. Four independent wells were used per condition.

For experiments involving murine organoids, *Nlrc5*$^{+/+}$ and *Nlrc5*$^{-/-}$ lines were grown for 3 days after splitting before treatment with or without recombinant murine IFNγ (Thermo Fisher #PMC4031) for 24 hours. Organoids were then harvested for RNA extraction. Two independent wells for each of two organoid lines generated from separate male mice per genotype were used.

NLRC5-mCherry PiggyBac plasmid construct and IEOs transfection

The human *NLRC5* cDNA (*myc-NLRC5*) was obtained from AddGene (#37509). PiggyBac (PB)-*mCherry* transposon backbone plasmid, the PB transposase and rtTA-hygromycin resistance (rtTA-HygRes) plasmids were a generous gift from B.K. Koo (Institute of Molecular Biotechnology, the Austrian Academy of Sciences). *myc-NLRC5* cDNA was obtained by PCR amplification of the *myc-NLRC5* plasmid using Phusion High-Fidelity DNA Polymerase (New England BioLabs) and cloned into PB-*mCherry,* after being digested using XhoI and NotI (New

England BioLabs). Primers were designed using Clontech's online InFusion Cloning Primer Design Tool (Table S4).

Cloning of *myc-NLRC5* cDNA into the PB-*mCherry* transposon plasmid was carried out using the InFusion HD Cloning Plus kit (Takara Bio). The PB NLRC5-mCherry plasmid or the PB-mCherry backbone plasmid was used in combination with PB transposase and rtTA-HygRes plasmids to transfect TI-derived IEOs. Transfection of IEOs was adapted from the protocol described by[6]. On day 7 after transfection, resistant clones were selected using hygromycin B (25–100 μg/ml). Successful transfection and overexpression were confirmed by fluorescence microscopy and qPCR following culture of PB NLRC5-mCherry lines and PB mCherry organoids in the presence of 1 μg/ml doxycycline hyclate or vehicle for two days.

CRISPR-mediated gene editing of human IEO

Guide RNAs were designed with a predicted On-target score >40, Off-target score >80 and a frameshifting score of >80. Three RNA guides were designed to target *NLRC5*, and knock-out (KO) efficiency was tested in HEK293T cells. The guide producing the greatest KO per gene was selected for implementation in organoids, and effectiveness of the knock-out was evaluated by PCR using the primers: Fw 5'-ACTCACCTATCTTCCCTGTCC-3' and Rev 5'-TCGGGCATTATGGGCTATGT-3'.

Tissue Freezing, Sectioning and RNAscope

Human mucosal biopsies undergoing RNAscope analysis were washed five times with PBS, incubated in 30% sucrose solution in PBS overnight at 4ºC and then 4% formaldehyde solution in PBS for 24-28 hours, before embedding in optimal cutting temperature compound (OCT, Fisher Scientific, UK). Samples were processed and RNAscope experiments performed at the Cellular Generation and Phenotyping facility at the Wellcome Sanger Institute as described previously (Supplementary Materials and Methods)[7].

Cryosections were cut at a thickness of 10 μm using a Leica CM3050 S cryostat and placed onto SuperFrost Plus slides (VWR). Staining was performed using RNAscope multiplex fluorescent reagent kit v2 assay (Advanced Cell Diagnostics, Bio-Techne) with an automated Leica BOND RX according to the manufacturer's instructions, including epitope retrieval by protease digestion with protease III for 15 minutes at room temperature and a heat-induced epitope retrieval (HIER) using epitope retrieval 2 (ER2) at 88°C for 5 minutes. RNAscope probe hybridization, immunohistochemistry (IHC) primary and secondary staining and channel development with Opal 520, Opal 570, and Opal 650 dyes (Akoya Biosciences) were then

performed. Atto 425 (Sigma-Aldrich) was also developed using TSA-biotin (TSA Plus Biotin Kit, Perkin-Elmer). Details of probes and antibodies used are included in Table S5.

IEOs Quantitative Imaging

IEOs were passaged as described above, plated in 96-well optical plates (Cell Carrier Ultra, Perkin Elmer) in 5µl of Matrigel per well, and growth for 5 days following treatment with or without 20 ng/ml IFNg (Thermo Fisher) for 48 hours. They were then fixed in 4% formaldehyde solution (Sigma-Aldrich) for 1 hour, permeabilised with 0.5% Triton X-100 (Thermo Fisher) in PBS for 20 mins and blocked with 2.5% normal horse serum for 30 minutes at room temperature. The staining was performed by using HLA-A,B,C-APC antibody (see Table S6), diluted 1:100 in 0.1% BSA/PBS and incubated with the IEOs overnight at 4ºC, followed by counterstaining with 1 ug/ml DAPI (Thermo Fisher) in PBS for 5 mins at room temperature. IEOs were washed 3 times with PBS between each step. Imaging was then performed with the Opera Phoenix high content imaging system (Perkin Elmer) by using the confocal mode with a 10x objective lens (NA = 0.3) and z stacks with a step size of 7.4 um. Camera ROI of 2160x2160 px with 2x2 binning was used. Image analysis and quantification were performed using Harmony High-Content Imaging and Analysis Software version 5.1 (Perkin Elmer). Image regions were identified using DAPI intensity and HLA-A,B,C levels were quantified by calculating image region mean APC intensity and averaging across each well. Experiments were performed in triplicate.

Murine Intestinal epithelial organoid culture, organoid peptide stimulations and co-culture with OTI T cells

Nlrc5$^{-/-}$ or *Nlrc5*$^{fl/f/l}$ mouse organoids were set up from two age and sex-matched mice per genotype, as previously described[8, 9].

Nlrc5$^{-/-}$ or *Nlrc5*$^{fl/f/l}$ organoids were cultured in 24-well plates for two to four days from splitting, then stimulated with or without 20 ng/ml IFNγ for 24 or 48 hours as indicated. Organoids were then digested with 500 µl TrypLE per well for up to 15 minutes at 37ºC, with vigorous pipetting at 5-minute intervals, until most of them had broken up. Cells were then transferred to 96 well V bottom plates and pulsed with 2 ug/ml $OVA_{257–264}$ (Bachem) peptide for 2 hours at 37ºC, before proceeding to immunostaining and analysis by flow cytometry.

To generate activated OTI T cells, spleens harvested from OT-I *RAG1*-deficient mice were homogenized through a 70 µm filter and stimulated at 0.5 x 10^6 cells/ml with 10 nM $OVA_{257–}$

$_{264}$ peptide (AnaSpec) for three days in OTI T cell medium (RPMI-1640 plus 10% FBS, 2 mM L-glutamine, 50 U/ml penicillin/streptomycin, 50 mM β-mercaptoethanol (βME, Thermo Fisher) and 20 ng/ml recombinant murine IL-2 (mIL-2, Peprotech). On day 3, cells were washed once and seeded in fresh medium supplemented with mIL-2 at 0.5 x 10^6 cells/ml every 48 hours. CTLs were considered ready for use on day 6-7 after initial stimulation.

For organoid-OTI T cell co-cultures, mouse organoids were loaded with $OVA_{257–264}$ peptide by digesting with TrypLE and stimulated with the peptide, pulsing for 2 hours as described above. Cells were then washed three times in PBS and cells from a single well of a 48 well plate of organoids were cultured with 200 x 10^5 activated T cells per well in 200 µl T cell medium. Co-cultures were incubated for 4 hours at 37ºC, 5% CO_2, before processing for RNA extraction. Cocultures were performed using two independent wells for each of the two organoid lines derived from separate C57B/6 male mice.

The dextran sulphate sodium colitis model

For all *in vivo* colitis experiments, eight weeks-of-age, female mice were used. Experimental groups were co-housed for three weeks prior to use to acclimatise mice and minimise confounding. Colitis was induced by supplementation of 2% w/v 36,000-50,000 MW dextran sulphate sodium (DSS, MP Biomedicals) in drinking water for six days. Body weight was monitored every day to define colitis severity. A threshold of 20% weight loss was set as a moderate severity limit, however, no mice exceeded this limit. On day 14 after the first administration of DSS, mice were euthanised and tissues harvested. Spleens, MLNs, and colons were weighed, and colon lengths measured from cecum to rectum. Tissues were then processed for flow cytometric analysis. Experimentalists were blinded to experimental groups during data collection.

Histological assessment of murine intestinal samples: Sections of the mouse distal small intestine and colon were fixed in 10% neutral-buffered formalin and processed by the Wellcome-MRC Cambridge Stem Cell Institute histology core facility. Paraffin embedding, sectioning, and hematoxylin and eosin (H&E) staining were conducted following established protocols, with images captured using standard light microscopy.

Inflammatory scoring was conducted based on a combined system developed by Erben *et al*.[10]. Total score range was 0-8 (small bowel) and 0-6 (colon), consisting of combined ‘inflammatory cell infiltrate’ and ‘mucosal architecture’ scores (each 0-4 and 0-3 for small bowel and colon respectively). Scoring procedures were carried out by a trained histopathologist, ensuring impartiality by maintaining blindness to the experimental groups.

Flow cytometry

Mouse and human organoids were dissociated into single cells by incubation at 37ºC in TrypLE for 10-15 minutes to perform flow cytometry. Cells were washed in FACS buffer (PBS + 5 mM EDTA + 1% BSA) and stained for 30 minutes at 4ºC in the dark with an antibody mix used at 1:200 dilution in FACS buffer (antibodies listed in Table S6). Cells were then washed with PBS, stained for dead cells using Zombie aqua viability dye (Biolegend) at 1:300 dilution in PBS for 20 minutes at room temperature in darkness and washed again with PBS. Cells were then analysed immediately using a Fortessa flow cytometer (Becton Dickinson).

DATA ANALYSES

DNA Methylation Data and analysis

Genome-wide DNA methylation was profiled using either the Illumina Infinium Human Methylation 450 BeadChip, or the Illumina EPIC platform (Illumina, Cambridge, UK) [11],[12]. DNAm data was processed using the *minfi* package and normalised based on control probes on each array using functional normalisation[13]. Samples were removed if they had high average detection p-values (>0.05) across all probes. Batch effect was removed by using *ComBat[14]* and probes were filtered for quality leaving 799,922 and 799,049 CpGs in the IEO and IEC EPIC array samples respectively, and 414,293 CpGs in the IEC 450K array samples[11],[12]. As the IEC samples were assayed on both Illumina EPIC and 450K arrays, only the 384,394 CpGs measured on both platforms were used for differential DNAm analysis in this cohort (Cohort 2, Table S2). Of a total of 142 at diagnosis IEC samples, 82 were present on the 450K array. All IEO samples were assayed on the Illumina EPIC array (Cohort 1, Table S2). *Accession Numbers: E-MTAB-12841, E-MTAB-5463.*

Previously published genome-wide DNAm profiles of primary purified IEC samples[5] were subjected to in-silico correction for potential cellular contamination. Briefly, cell composition was estimated using existing/public datasets from the Gene Expression Omnibus (GEO)[15] and ArrayExpress[16] then IEC DNAm values were adjusted for intraepithelial lymphocyte proportion[17].

Differential DNA Co-Methylation

Epigenome-wide weighted gene co-expression network analysis (WGCNA)[18] was performed on both TI and SC IEO samples (Cohort 1). The top 10% variant probes, identified across all samples (TI: n=127, SC: n=131), were retained after filtering for low variance beta values, resulting in 79,757 CpGs. Co-methylation network analysis was carried out using the WGCNA R package (v.1.72.). An appropriate soft-thresholding power (β = 8) was selected to fit a scale-free network. Signed topological overlap matrix (TOM) was calculated to measure the pairwise Pearson's correlation between all CpGs. The resulting hierarchical clustering dendrogram was generated, followed by module detection with a minimum module size 100 and a minimum cut height of 0.995, using the Dynamic Tree Cut algorithm. Subsequently, the first principal components of each module and their correlation with disease traits were calculated to identify the set of co-methylated CpGs strongly associated with CD.

Pathway Enrichment

To explore pathways related to differential methylation with IBD, the enrichment of gene ontology (GO)[19] groups in the genes adjacent to hypomethylated CpGs was performed using *fgsea* (v1.24.0)[20] and GO biological processes as pathways. Only GO groups between 15 and 400 genes were used.

DNAm and Expression Correlation

Correlation between DNAm and expression was explored by calculating Spearman's correlations and compared to random sets on comparable CpG gene pairs**.** Ensembl gene IDs with the highest mean count in the IEC data were used, which left 29,411 unique genes. Correlations between DNAm at CpGs adjacent to a gene and that gene's expression (26,963 correlations; 2-1019 CpGs/gene, mean 28 CpG per gene; 1-21 gene/CpG, mean 1.1) were calculated for all CpG-gene pairs.

Expression Array Data

Publicly available TI (n=78) and colon (n=116) biopsy expression array data from adult patients profiled on the Affymetrix Human Gene 1.0 ST expression array were downloaded from the GEO (Cohort 7, n=75 CD, 97 UC and 22 non-IBD controls, Table S2)[21]. *Accession Number: GSE75214*. Raw signal intensity data were normalised using the Robust Multi-array Average (RMA) method using *affy v1.70.0* and quality control was performed using

arrayQualityMetrics v3.48.0[22, 23, 24]. The normalised dataset was annotated using the *pd.hugene.1.0.st.v1 v3.14.1* annotation package[25].

Bulk RNA-seq Data

Paediatric purified epithelium from terminal ileum (TI) and sigmoid colon (SC), Cohort 2 (n = 20 CD, 22 UC and 22 non-IBD controls, Table S2), were transcriptionally profiled using RNA-seq by University of Kiel, Germany using an established pipeline as described previously. *Accession Number: E-MTAB-5464*.

Unmodified human, PB NLRC5-mCherry and *Nlrc5*$^{+/+}$ and *Nlrc5*$^{-/-}$ mouse organoid lines treated with inflammatory cytokines as described above were transcriptionally profiled at Cambridge Genomic Services (Dept. Pathology, Tennis Court Road, University of Cambridge, Cambridge). *Accession Numbers: E-MTAB-11548; NLRC5-mCherry and mouse data available on request*.

In addition, publicly available TI biopsy RNA-seq data from 322 paediatric patients were downloaded from the GEO using SRA Explorer (Cohort 6, n = 218 CD, 62 UC and 42 non-IBD controls, Table S2)[26]. *Accession Number: GSE57945*.

Raw FastQ files for all bulk RNA-seq samples were first pre-processed using *BBMap v38.26[27]* to remove adapters and low-quality sequence. *FastQC v0.11.9* was used to ascertain data quality before and after trimming and filtering[28]. Reads were next aligned to the human or mouse genome (Ensembl GRCh38 and GRCm38, release 99) using *STAR v2.5.0* and indexed using *SamTools v1.11*. Finally, raw counts were extracted using *featureCounts*, (*subread* package, *v2.0.3*) and assessed for quality using the *NOISeq v2.36.0* and *DESeq2 v1.32.0*, R packages[29, 30, 31, 32, 33].

Average MHC-I Expression Score

An average MHC-I gene expression score comprising the genes *NLRC5*, *TAP1/2*, *PSMB8/9*, *HLA-A/-B/-C/-E/-F/-G*, *IRF1* and *B2M* was calculated, subtracted by the aggregated expression of 100 control gene sets. All analysed genes were binned based on average expression, and control genes randomly selected from each bin.

Single cell RNA-seq data analyses

Raw sequencing reads were aligned with CellRanger software (V302; 10x Genomics) to the reference human transcriptome (GRCh38), then the resulting unique molecular identifier (UMI) counts were processed using *Seurat* 4.1.0[34, 35]. Cells with fewer than 500, more than

6,000 detected genes, or with a mitochondrial transcript proportion higher than 10% were excluded. In addition, a total of 256 doublets were detected using *Scrublet[36]*.

For the whole-biopsy samples (Cohorts 3 and 4, Table S2), cell annotation had been expertly assigned previously. For IEOs (Cohort 5, Table S2), cell type labels were assigned within each treatment group to preserve differences in gene expression arising from the proinflammatory cytokines treatment. Cell type labels were transferred from the annotated whole-biopsy samples to organoid cell clusters and confirmed with marker genes (Table S7).

For differential gene expression (DE), raw count data was used, log-transformed, scaled by the total counts in a given cell, and multiplied by 10,000. Differential expression of individual genes between treatments in organoids and diagnosis in whole-biopsies was performed within each identified cell type using the *FindMarkers* function in Seurat, with MAST including number of genes detectable in a cell (nFeature_RNA) as a latent variable.[37] Unless otherwise indicated, all reported DE coefficients and q-values correspond to the combined hurdle p-value from the likelihood-ratio-test.

To calculate overall MHC-I and crypt-villus axis scores for each cell, *addModuleScore* from Seurat was used on raw, log-transformed, scaled count data[35]. The scores are equivalent to a mean expression of the given genes and adjusted for the mean expression of each gene across all cells, in order to prevent highly expressed genes from overly contributing to the score. *Accession Numbers: E-MTAB-8901; SCP1884*.

Differential DNA Methylation

Epigenome-wide association tests were performed in both the IEO and IEC DNAm cohorts (Cohorts 1 and 2). In the IEC samples, covariates for histological based inflammation measure (binary: inflamed or not), sex and age were incorporated with linear models using *limma*. The difference in group mean betas (delta beta, db) was also calculated as a measure of effect size. As the IEOs are a pure population of epithelial cells with no immune cell component, inflammation status was irrelevant and was not included as a covariate in the model. Age and sex were included as covariates (Tables S8, S9, S10).

Statistical Univariate Analysis of DNA Methylation and Biomarker Selection

For CD severity prognostic methylation biomarkers, TI CD (n=55) IEOs (Cohort 1) samples were classified into two groups, severe CD and mild/moderate CD. For each individual CpG, two groups of samples were used to generate two empirical distributions, and their difference was assessed by 2-sample Anderson-Darling test[38]. The CpGs with smaller p-values have

more distinct distributions between the two groups, so they were selected as informative features for the prediction of whether a sample is from one group or the other. Using the method described above, 78 CpGs candidates for diagnosis and 72 CpGs candidates for prognosis were selected. Taking the intersection with significant CpGs from WGCNA DNA Co-Methylation network analysis, 53 CpGs were selected as CD diagnostic biomarker (Table S9, p-value for intersection < 1e-5), while 28 CpGs were selected as CD severity prognostic biomarker (Table S10, p-value for intersection < 1e-5).

Machine Learning Predictors for CD Prognosis and Diagnosis

To develop CD severity prognostic prediction model, preliminary model selection using default hyperparameters were carried out to test different machine learning classification algorithms. With 628 MHC-I CpGs (Table S8) or 28 prognostic CpGs (Table S10) of TI (n=127) IEOs (Cohort 1), logistic regression (AUC 0.71/0.72) performs consistently better than support vector machine (AUC 0.71/0.62)[39], random forest (AUC 0.67/0.66)[40], XGBoost (AUC 0.64/0.62)[41] and perceptron (AUC 0.71/0.62)[42]. Therefore, logistic regression classifier was selected for the subsequent analysis. Healthy and disease control (UC) samples were labelled as non-severe CD and used in training data to provide more context for classifiers, while only CD samples were used in testing. Repeated cross-validation was chosen as the resampling method to estimate model performance. Cross-validations were stratified such that each split has the same number of control, mild CD, moderate CD and severe CD samples. Cross-validation split number was chosen to be 5. Other split numbers were tested and produced similar results. Evaluation metrics for predictors were chosen to be ROC curve and corresponding AUC scores, and their estimated population means obtained from repeated cross-validation were used for comparison. To illustrate the prognostic value of 628 MHC-I CpGs (Table S8), permutation analysis was done by randomly selecting 628 CpGs from all measured CpGs and running the aforementioned prediction procedure. By AUC scores, predictors with 628 MHC-I CpGs as input perform better than 95.8% (95% CI size < 2%) of randomly selected CpG sets. The median performer among randomly selected CpG sets was included in the ROC plot (Figure 7Ciii). Machine learning classification algorithms, cross-validation and evaluation metrics were from scikit-learn library[43]. The same procedure was also implemented to develop CD diagnostic prediction model.

Risk Score System and Patient Stratification

The prognostic risk score system was developed by further processing the output of machine learning prediction model. Machine learning binary classifier operates by generating probability predictions for each sample and comparing it to a fixed classification threshold between 0 and 1 to make a decision. Each logistic regression classifier from repeated cross validation generated a probability prediction of the binary outcome (severe/non-severe) for each sample in its testing set. Since each sample appears in the testing set of multiple classifiers, the cross-sectional mean of probability predictions for each sample was taken to generate a stable estimate for mean risk of developing severe CD. To ensure the consistency of cross-sectional operation between different classifiers, a coherent methodology to choose classification threshold and adjust probability predictions accordingly for each classifier was needed. After a comparison of methods to select the optimal threshold, the method of always choosing the top left corner of ROC (minimising the top left index, (1 − TPR)^2+FPR^2) was selected[44]. To generate reliable risk scores, cross-validation was repeated until risk score of every sample converged (size of 95% CI of mean < 0.01). The same procedure was also implemented to develop CD diagnostic risk score system. To demonstrate the risk score system's potential for prognostic stratification, two cut-offs (risk>0.65 and risk<0.33) were chosen to detect high and low risk patients respectively and were represented on Fagan's Nomogram (Figure 7Civ). 60% (95% CI: 46%-73%) of CD patients are expected to be stratified by one of the two cut-offs.

Data availability

DNA Methylation data are available at ArrayExpress16 with Accession Numbers: E-MTAB-12841, E-MTAB-5463, Bulk RNA-seq data are available at ArrayExpress with Accession numbers: E-MTAB-5464., E-MTAB-11548. NLRC5-mCherry and mouse data are available on request. Single-cell RNA-seq data are available at ArrayExpress and SCP (singlecell.broadinstitute.org/single_cell) with Accession Numbers: E-MTAB-8901; SCP1884. All codes generated in this study are deposited and available on GitHub: https://github.com/ZilbauerLab/NLRC5_MHCI.

Full dataset repository: https://www.ebi.ac.uk/biostudies/arrayexpress/studies/E-MTAB-12841?key=8aa7741a-673a-480f-a184-a73a7d9295b5

Code Availability

All codes generated in this study are deposited and available on GitHub: https://github.com/ZilbauerLab/NLRC5_MHCI.

References

1 Howell KJ, Kraiczy J, Nayak KM, Gasparetto M, Ross A, Lee C*, et al.* DNA Methylation and Transcription Patterns in Intestinal Epithelial Cells From Pediatric Patients With Inflammatory Bowel Diseases Differentiate Disease Subtypes and Associate With Outcome. Gastroenterology 2018;**154**:585-98.
2 Kraiczy J, Nayak KM, Howell KJ, Ross A, Forbester J, Salvestrini C*, et al.* DNA methylation defines regional identity of human intestinal epithelial organoids and undergoes dynamic changes during development. Gut 2019;**68**:49-61.
3 Edgar RD, Perrone F, Foster AR, Payne F, Lewis S, Nayak KM*, et al.* Culture-Associated DNA Methylation Changes Impact on Cellular Function of Human Intestinal Organoids. Cell Mol Gastroenter 2022;**14**:1295-310.
4 Kraiczy J, Nayak KM, Howell KJ, Ross A, Forbester J, Salvestrini C*, et al.* DNA methylation defines regional identity of human intestinal epithelial organoids and undergoes dynamic changes during development. Gut 2019;**68**:49-61.
5 Edgar RD, Perrone F, Foster AR, Payne F, Lewis S, Nayak KM*, et al.* Culture-Associated DNA Methylation Changes Impact on Cellular Function of Human Intestinal Organoids. Cell Mol Gastroenterol Hepatol 2022;**14**:1295-310.
6 Fujii M, Matano M, Nanki K, Sato T. Efficient genetic engineering of human intestinal organoids using electroporation. Nature Protocols 2015;**10**:1474-85.
7 Elmentaite R, Ross ADB, Roberts K, James KR, Ortmann D, Gomes T*, et al.* Single-Cell Sequencing of Developing Human Gut Reveals Transcriptional Links to Childhood Crohn's Disease. Dev Cell 2020;**55**:771-+.
8 Sato T, Vries RG, Snippert HJ, van de Wetering M, Barker N, Stange DE*, et al.* Single Lgr5 stem cells build crypt-villus structures in vitro without a mesenchymal niche. Nature 2009;**459**:262-5.
9 Sato T, Stange DE, Ferrante M, Vries RG, Van Es JH, Van den Brink S*, et al.* Long-term expansion of epithelial organoids from human colon, adenoma, adenocarcinoma, and Barrett's epithelium. Gastroenterology 2011;**141**:1762-72.
10 Erben U, Loddenkemper C, Doerfel K, Spieckermann S, Haller D, Heimesaat MM*, et al.* A guide to histomorphological evaluation of intestinal inflammation in mouse models. Int J Clin Exp Patho 2014;**7**:4557-U27.
11 Pidsley R, Zotenko E, Peters TJ, Lawrence MG, Risbridger GP, Molloy P*, et al.* Critical evaluation of the Illumina MethylationEPIC BeadChip microarray for whole-genome DNA methylation profiling. Genome Biology 2016;**17**:208.
12 Price ME, Cotton AM, Lam LL, Farré P, Emberly E, Brown CJ*, et al.* Additional annotation enhances potential for biologically-relevant analysis of the Illumina Infinium HumanMethylation450 BeadChip array. Epigenetics Chromatin 2013;**6**:4.

13 Aryee MJ, Jaffe AE, Corrada-Bravo H, Ladd-Acosta C, Feinberg AP, Hansen KD*, et al.* Minfi: a flexible and comprehensive Bioconductor package for the analysis of Infinium DNA methylation microarrays. Bioinformatics 2014;**30**:1363-9.
14 Leek JT, Johnson WE, Parker HS, Jaffe AE, Storey JD. The sva package for removing batch effects and other unwanted variation in high-throughput experiments. Bioinformatics 2012;**28**:882-3.
15 Edgar R, Domrachev M, Lash AE. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. Nucleic Acids Res 2002;**30**:207-10.
16 Parkinson H, Kapushesky M, Shojatalab M, Abeygunawardena N, Coulson R, Farne A*, et al.* ArrayExpress—a public database of microarray experiments and gene expression profiles. Nucleic Acids Research 2006;**35**:D747-D50.
17 Houseman EA, Accomando WP, Koestler DC, Christensen BC, Marsit CJ, Nelson HH*, et al.* DNA methylation arrays as surrogate measures of cell mixture distribution. BMC Bioinformatics 2012;**13**:86.
18 Langfelder P, Horvath S. WGCNA: an R package for weighted correlation network analysis. BMC Bioinformatics 2008;**9**:559.
19 The Gene Ontology resource: enriching a GOld mine. Nucleic Acids Res 2021;**49**:D325-d34.
20 Korotkevich G, Sukhov V, Budin N, Shpak B, Artyomov MN, Sergushichev A. Fast gene set enrichment analysis. bioRxiv 2021:060012.
21 Vancamelbeke M, Vanuytsel T, Farré R, Verstockt S, Ferrante M, Van Assche G*, et al.* Genetic and Transcriptomic Bases of Intestinal Epithelial Barrier Dysfunction in Inflammatory Bowel Disease. Inflammatory Bowel Diseases 2017;**23**:1718-29.
22 Gautier L, Cope L, Bolstad BM, Irizarry RA. affy--analysis of Affymetrix GeneChip data at the probe level. Bioinformatics 2004;**20**:307-15.
23 Irizarry RA, Hobbs B, Collin F, Beazer-Barclay YD, Antonellis KJ, Scherf U*, et al.* Exploration, normalization, and summaries of high density oligonucleotide array probe level data. Biostatistics 2003;**4**:249-64.
24 Kauffmann A, Gentleman R, Huber W. arrayQualityMetrics—a bioconductor package for quality assessment of microarray data. Bioinformatics 2008;**25**:415-6.
25 Carvalho B. pd. hugene. 1.0. st. v1: Platform Design Info for Affymetrix HuGene-1_0-st-v1. R package 2015.
26 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.
27 B. B-B.
28 de Sena Brandine G, Smith AD. Falco: high-speed FastQC emulation for quality control of sequencing data. F1000Res 2019;**8**:1874.
29 Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S*, et al.* STAR: ultrafast universal RNA-seq aligner. Bioinformatics 2013;**29**:15-21.
30 Liao Y, Smyth GK, Shi W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. Bioinformatics 2014;**30**:923-30.
31 Tarazona S, Furió-Tarí P, Turrà D, Pietro AD, Nueda MJ, Ferrer A*, et al.* Data quality aware analysis of differential expression in RNA-seq with NOISeq R/Bioc package. Nucleic Acids Research 2015;**43**:e140-e.
32 Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol 2014;**15**:550.

33 Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N*, et al.* The Sequence Alignment/Map format and SAMtools. Bioinformatics 2009;**25**:2078-9.
34 Hao Y, Hao S, Andersen-Nissen E, Mauck WM, 3rd, Zheng S, Butler A*, et al.* Integrated analysis of multimodal single-cell data. Cell 2021;**184**:3573-87.e29.
35 Tirosh I, Izar B, Prakadan SM, Wadsworth MH, Treacy D, Trombetta JJ*, et al.* Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-seq. Science 2016;**352**:189-96.
36 Wolock SL, Lopez R, Klein AM. Scrublet: Computational Identification of Cell Doublets in Single-Cell Transcriptomic Data. Cell Syst 2019;**8**:281-91.e9.
37 Finak G, McDavid A, Yajima M, Deng J, Gersuk V, Shalek AK*, et al.* MAST: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data. Genome Biol 2015;**16**:278.
38 Scholz FW, Stephens MA. K-Sample Anderson-Darling Tests. Journal of the American Statistical Association 1987;**82**:918-24.
39 Cortes C, Vapnik V. Support-vector networks. Machine Learning 1995;**20**:273-97.
40 Breiman L. Random Forests. Machine Learning 2001;**45**:5-32.
41 Chen T, Guestrin C. XGBoost: A Scalable Tree Boosting System. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. San Francisco, California, USA: Association for Computing Machinery, 2016:785–94.
42 Rosenblatt F. The perceptron: A probabilistic model for information storage and organization in the brain. Psychological Review 1958;**65**:386-408.
43 Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O*, et al.* Scikit-learn: Machine Learning in Python. J Mach Learn Res 2011;**12**:2825–30.
44 De Borre M, Che H, Yu Q, Lannoo L, De Ridder K, Vancoillie L*, et al.* Cell-free DNA methylome analysis for early preeclampsia prediction. Nat Med 2023;**29**:2206-15.