nature communications

Article

https://doi.org/10.1038/s41467-025-62533-z

Combining mucosal microbiome and host multi-omics data shows prognostic potential in paediatric ulcerative colitis

Received: 7 February 2025

Accepted: 23 July 2025

Published online: 04 August 2025

Check for updates

Maria Kulecka[1,2], Jill O'Sullivan[1,2,3], Rachel Fitzgerald[1,2], Ana Velikonja[1,2], Chloe E. Huseyin[1,2], Emilio J. Laserna-Mendieta[1,2], Patricia Ruiz-Limón[1,4], Julia Eckenberger[1,5], Miriam Vidal-Marín[1], Bastian-Alexander Truppel[1], Raminder Singh[1], Sandhia Naik[6], Nicholas M. Croft[7], Andriy Temko[8], Aldert Zomer[9], John MacSharry[1,2], Silvia Melgar[1], Protima Deb[6], Ian R. Sanderson[7] & Marcus J. Claesson[1,2]

Current first-line treatments of paediatric ulcerative colitis (UC) maintain a 6-month remission in only half of the patients. Relapse prediction at diagnosis could enable earlier introduction of immunosuppressants. We collected intestinal biopsies from 56 treatment-naïve children, combining mucosal quantitative microbial profiling with host epigenomics, transcriptomics, genotyping, and in vitro and in vivo experiments on selected bacteria. Baseline bacterial diversity is lower in relapsing children, who have fewer butyrate producers but more oral-associated bacteria, whereof *Veillonella parvula* induces inflammation in epithelial cell lines and IL10$^{-/-}$ mice. Microbiota has the strongest association with future relapse, followed by host epigenome and transcriptome. Interferon gamma signalling is also linked to relapse-associated bacteria. Relapse-prediction using separate omics data is outperformed by a robust machine learning approach combining microbiomes and epigenomes. In summary, host-microbe data have prognostic potential in paediatric UC. Our translational findings also suggest that pro-inflammatory oral-associated colonizers can exploit the reduced colonic bacterial diversity of relapsing children.

Paediatric ulcerative colitis (UC) is an immune-mediated disease of complex aetiology, with onset in children up to 17 years of age[1], characterized by mucosal inflammation of the colon, extending from the rectum[2]. Paediatric disease is often more severe than its adult counterpart[3,4] and is a major public health concern due to its increased mortality rates[5] and cancer risks[6] throughout lifetime. The incidence is increasing worldwide, with Western European countries presenting with one of the highest incidence rates[7]. Etiopathogenesis of UC is multi-factorial and complex, with paediatric UC having a stronger genetic component than adult disease[8,9]. Implicated susceptibility

[1]APC Microbiome Ireland, University College Cork, Cork, Ireland. [2]School of Microbiology, University College Cork, Cork, Ireland. [3]Research Ireland Centre for Research Training in Genomics Data Science, University of Galway, Galway, Ireland. [4]Department of Endocrinology and Nutrition, Virgen de la Victoria University Hospital, The Biomedical Research Institute of Malaga and Platform in Nanomedicine, Malaga, Spain. [5]College of Health and Medicine, University College Cork, Cork, Ireland. [6]Paediatric Gastroenterology, Barts Health NHS Trust, London, UK. [7]Centre for Immunobiology, Blizard Institute, Queen Mary University of London, London, UK. [8]Department of Electrical and Electronic Engineering, University College Cork, Cork, Ireland. [9]Department of Biomolecular Health Sciences (Infectious Diseases and Immunology), Faculty of Veterinary Medicine, Utrecht University, Utrecht, Netherlands.
e-mail: M.Claesson@ucc.ie

genes are involved in barrier function, adaptive immunity, and antigen presentation[10], suggesting impaired immune response to commensal and pathogenic bacteria alike. Consequently, gut microbial signatures have been suggested in a contributing role, both for paediatric[11,12] and adult[13,14] disease, as well as having prognostic potential[15] (reviewed in[16] for paediatric disease).

Children with UC are diagnosed after colonoscopy and biopsy, after which remission is induced with aminosalicylates (5-ASA) drugs and (usually) corticosteroids. A small number do not respond and will require ciclosporin, anti-TNF drugs, or colectomy. While remission is successfully achieved in most children, around half of them will relapse within 6 months[17]. A further remission is induced with a second course of corticosteroids with the addition of a thiopurine as an immunosuppressant[18]. Identification of children who are likely to relapse within 6 months of diagnosis would be advantageous, as they could be treated with immunosuppressants or biologics from the onset[19].

The available biomarkers for the aminosalicylates and corticosteroid treatment outcomes in paediatric UC are unfortunately limited. Various strategies have been adopted, including clinical features at diagnosis[20,21], carriage of single-nucleotide polymorphisms[22], and CD8 gene transcription in circulating CD8+ T-cells[23,24]. Other clinical features previously associated with sustained corticosteroid-free remission (52 weeks) include low baseline severity, baseline haemoglobin, week 4 clinical remission[25], and decrease in faecal calprotectin by week 12[26]. While these approaches show some promise, they have significant problems. First, using clinical criteria, the best prognostic index is the disease activity 4–12 weeks after treatment and not a marker before it[21,25]; second, a large-scale gene-association study failed to link most clinical subphenotypes of IBD to known genetic loci[27]; and third, isolating CD8 + T-cells from children requires significant amounts of blood from venipuncture to obtain sufficient lymphocytes. We suggest that it is likely that the environment of the intestine, to which immune cells respond, will have a determining effect, and this forms the basis of a new approach to a long-standing goal.

Biomarkers based on high-throughput sequencing technologies ("omics") of rectal mucosa have also been proposed and include antimicrobial peptide gene signature and bacteria from the *Ruminococcaceae* family and *Sutterella* genus[25]. Here, we applied a multi-omics approach on both host and microbial features from intestinal biopsies using combinatory machine learning (ML) modelling. We found that the mucosa-related microbiota (including some of oral origin) at UC diagnosis and their interaction with the mucosal transcriptome, and to a lesser degree the epigenome, predict the therapeutic outcome (at 26 weeks post-diagnosis following treatment with corticosteroids and aminosalicylates).

Results

Mucosal microbiome diversity and composition at diagnosis are associated with future relapse

We recruited 56 children between 2 and 16 years who were diagnosed as having UC through colonoscopy (see patient characteristics in Table 1). At diagnostic colonoscopy, 201 biopsies were collected from the terminal ileum (TI), ascending colon (AC), distal colon (DC), and rectum (R).

We sequenced the 16S rRNA V3V4 region to determine mucosal microbiota diversity and composition. After filtering and chimera-elimination, 156,565 reads per sample (range: 12,070–646,470, median: 137,686) generated 5196 ASVs, whereof 4464 ASVs were classified as bacteria (range: 29–368, median: 141). 16S rRNA amplicon sequencing was normalized to the total 16S rRNA gene copy number to better approximate abundance[28]. Principal coordinate analysis (PCoA) of Bray–Curtis distances, performed on a prevalence-filtered subset of 292 ASVs present in more than 10% samples, revealed significant shifts in both the first and third principal coordinates associated with future relapse (PCo1, PCo3, respectively; pooled samples, Wilcoxon test $p < 0.05$, Fig. 1A; DC samples, Fig. 1B). At the time of diagnosis, PUCAI was also associated with the microbiota PCo1 and PCo3 (Fig. 1C). However, no significant association with future relapse was detected by assessing the severity of the disease (as measured by the PUCAI) at the time of diagnosis (Fig. 1D). Analysis of pooled samples across all sites showed a significantly lower microbiota diversity (Shannon) and richness (Chao1) in subjects who later relapse (Fig. 1E). These differences remained significant when separated into descending colon and rectum (Supplementary Fig. 1).

We used a negative binomial model to identify 23 differential ASVs (Supplementary Data 1), with most taxa (16) depleted in the

Table 1 | Cohort characteristics, including the full cohort as well as split by single omics

	Full cohort, *N* = 56		Epigenetics, *N* = 55, Biopsies = 168		Microbiome, *N* = 48, Biopsies = 159		Transcriptomics, *N* = 53, Biopsies = 173	
	Relapse *N* = 28	Remission *N* = 28	Relapse *N* = 28	Remission *N* = 27	Relapse *N* = 23	Remission *N* = 25	Relapse *N* = 27	Remission *N* = 26
Age (months)								
Median (IQR)	144 (111, 158)	149 (119, 173)	144 (111, 158)	150 (118, 174)	140 (111, 157)	150 (126, 174)	148 (112, 158)	149 (121, 174)
Range	24–189	63–195	24–189	63–195	24–189	63–195	24–189	63–195
Sex								
F	13 (46%)	14 (50%)	13 (46%)	13 (48%)	12 (52%)	12 (48%)	13 (48%)	14 (54%)
M	15 (54%)	14 (50%)	15 (54%)	14 (52%)	11 (48%)	13 (52%)	14 (52%)	12 (46%)
PUCAI at diagnosis								
Median (IQR)	50 (35, 70)	55 (39, 70)	50 (35, 70)	55 (38, 70)	50 (35, 70)	55 (35, 65)	50 (35, 70)	55 (38, 70)
Range	20–80	20–85	20–80	20–85	20–80	20–85	20–80	20–85
Antibiotics use	6 (21%)	4 (14%)	6 (21%)	4 (15%)	3 (13%)	3 (12%)	6 (22%)	4 (15%)
Biopsies								
R			28	26	21	21	27	25
DC			13	3	22	18	18	10
AC			24	26	19	18	24	24
TI			23	25	18	22	21	24

R rectum, *DC* descending colon, *AC* ascending colon, *TI* terminal ileum.

https://doi.org/10.1038/s41467-025-62533-z

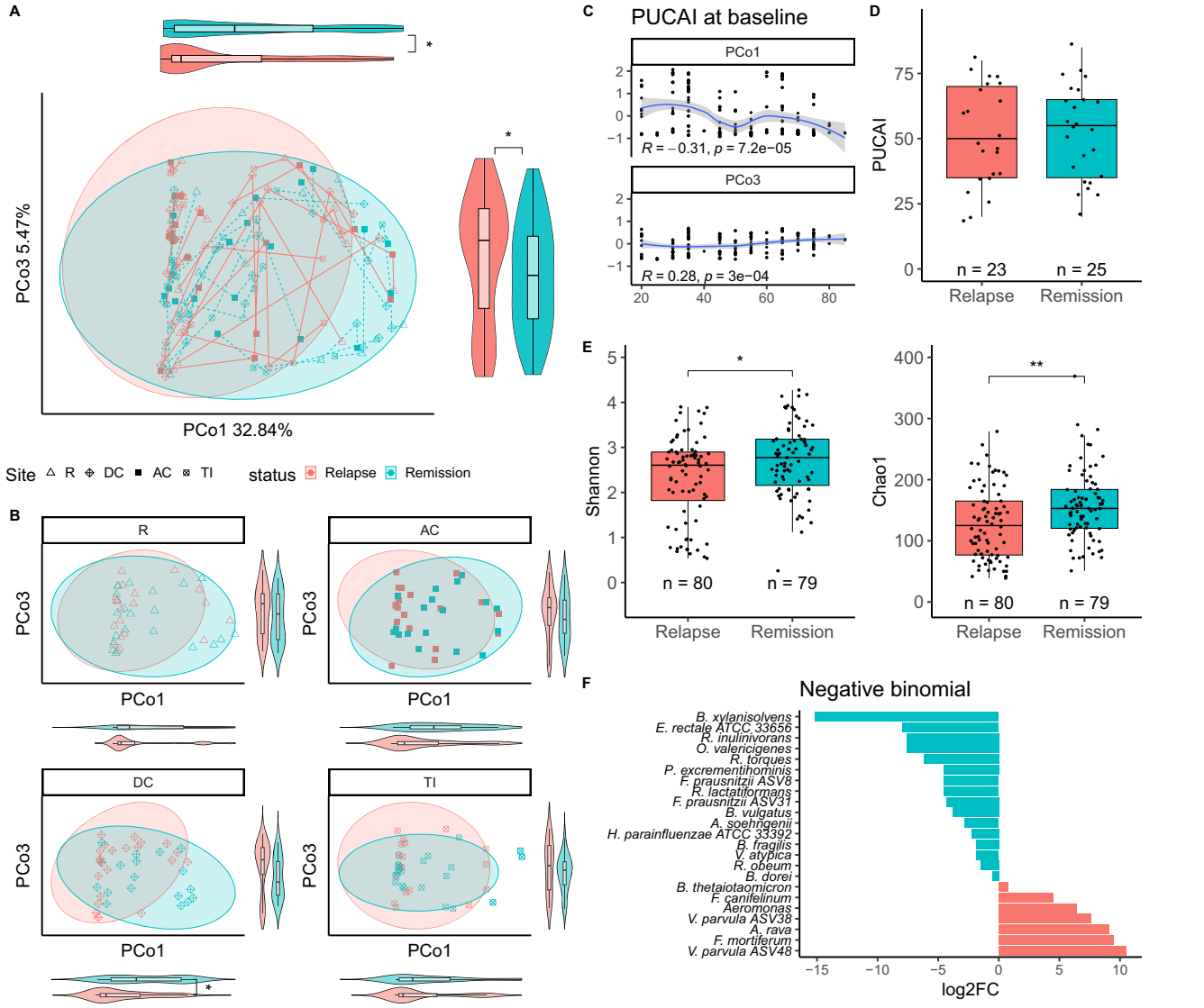

Fig. 1 | Microbiome analysis based on 16S V3-V4 data, corrected for absolute abundance with 16S qPCR. A Principal coordinate (PCo) analysis, based on Bray–Curtis distances. Depicted are coordinates significantly associated with the patient's status at 6 months (Wilcoxon's test, *n* Relapse = 80, *n* Remission = 79). **B** Principal coordinates analysis split by biopsy site. Relapse/remission numbers for sites are as follows: R:21/21, DC:22/18, AC:19/18, TI:18/22. **C** Associations of PUCAI at baseline with 1st and 3rd PCo, as measured by Spearman's correlation coefficient. For the smoothing function, a 95% confidence interval is used. **D** Differences of baseline PUCAI between relapse and remission cohorts. **E** Differences between relapse and remission cohorts (on pooled samples) in taxa diversity and richness, measured by Shannon and Chao1 indices, respectively. **F** Differentially abundant taxa between relapse and remission cohorts (pooled samples, mixed effects models) as indicated by zero-inflated negative binomial models. Significance codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1. All tests were two-sided, with no multiple testing correction. For boxplots, the lower and upper hinges correspond to the first and third quartiles, while the upper whisker extends from the hinge to the largest value no further than 1.5 × IQR from the hinge (IQR inter-quartile range). The lower whisker extends from the hinge to the smallest value at most 1.5 × IQR of the hinge. The middle line denotes the median.

relapse group. These include known butyrate producers (*F. prausnitzii*, *E. rectale*, and *R. inulinivorans*[29]) and mucin degraders (*B. fragilis*, *B. thetaiotaomicron*, *B. vulgatus*, and *R. torques*[30]. Seven ASVs were over-represented in the relapse group, including *V. parvula, Alloprevotella rava*, and *Fusobacterium canifelinum* (Fig. 1F). As antibiotic use is postulated as a risk factor for IBD development[31–33], it was not an exclusion criterion in our study. However, since it might affect microbial composition in the long term, we also performed a separate analysis with patients exposed to antibiotics excluded (141 samples from 42 children). This included previously mentioned butyrate producers (*F. prausnitzii* and *E. rectale*), mucin degraders (*B. fragilis*), as well as *V. parvula, A. rava*, and *F. canifelinum* (Supplementary Data 2).

Relapsing patients are over-represented in the *Veillonellaceae*-enriched microbial cluster

To highlight more granular microbiota groupings, we identified nine sample clusters (see Methods), where the largest contained 36 samples (cluster 1) and the smallest contained 10 samples (cluster 9). Some overlap is visible between clusters, however, this is not unexpected due to the possible spectral nature of human gut microbial communities[34]. Two clusters had statistically significant differences in the proportion of future relapse samples in comparison to all other clusters combined (47%). Cluster 2 had a significantly higher proportion of future-relapse samples (74%, 95% CI: 51–89%), while cluster 8 contained only future-remission samples (0% of relapse samples, 95% CI: 0–32%). For cluster 2, families *Veillonellaceae* and *Aerococcaceae*

were significant indicators, and it also shared indicators *Enterobacteriaceae* and *Clostridiaceae* with cluster 5. For cluster 8, only *Acidaminococcaceae* was a significant indicator. At the more granular ASV level, *Escherichia–Shigella*, *Aggregatibacter segnis*, and *Veilonella dispar* were indicators for cluster 2, while *Prevotella copri* and unclassified *Lachnospiraceae* were indicators for cluster 8. Cluster 2 was also characterized with significantly lower species diversity (mean Shannon: 1.63, 95% CI: 1.16–2.11) and richness (mean Chao1: 115, 95% CI: 99–132) than the overall cohort mean, whereas cluster 8 had significantly higher species diversity than the cohort mean (mean Shannon index 3.26, 95% CI: 2.95–3.56). Cluster 2 participants were significantly younger than the full cohort (mean age: 128 months, 95% CI: 120–135), more likely to be female (74%, 95% CI: 51–89%), and less likely to be born vaginally (35%, 95% CI: 17–57%). Conversely, participants from cluster 8 were significantly older than the average of all other clusters (mean age: 156 months, 95% CI: 144–169). Interestingly, both groups had significantly higher PUCAI at the time of diagnosis than the cohort mean (Fig. 2 and Supplementary Data 3). We also performed clustering for each biopsy site separately, however, none of the clusters had any association with future patient status (Supplementary Fig. 2 and Supplementary Data 3).

V. parvula induces pro-inflammatory responses both in vitro and in vivo, which cannot be counteracted by UC therapeutics

To examine the potentially detrimental influence on the host and the effect of UC-associated therapies, the inflammatory effect of the significant relapse-associated bacteria was studied using bacterial conditioned media (i) co-cultured with intestinal epithelial cell lines and (ii) whole bacteria was orally gavaged to IL10$^{-/-}$ mice. Two *Veillonella* genus species were chosen: *V. dispar*, as it was an indicator species in a relapse-related bacterial cluster, and *V. parvula*, as it was significantly increased in the relapse group. *V. dispar* was chosen over other indicator species for its cluster because: (1) *Veillonellaceae* as a family were also identified as indicators in this cluster, (2) it showed higher abundance in the gut compared to *A. segnis*, and (3) for *Escherichia–Shigella*, a strain-level analysis would likely be necessary to identify a suitable strain—which was not feasible with our 16S rRNA data. Co-culture (24 h) of epithelial cells with *V. parvula* conditioned media significantly induced secretion of IL-8 and NFκB-activation, but not by *V. dispar* (Fig. 3A, C, D). While these species activated the IRF-pathway, this was not associated with CXCL10 secretion (under detection level of 15.6 pg/mL of ELISA assay) or reduced by the UC-associated JAK-inhibitor Tofacitinib, when compared to the control polydA:T (Fig. 3B). Similarly, treatment with Tofacitinib and methylprednisolone failed to reduce IL-8 secretion and NFκB-activation induced by these species (Fig. 3A, C, D). In contrast, Sulfasalazine (component of aminosalicylates) significantly reduced *V. parvula*-induced IL-8 and NFκB activation (Fig. 3C, D). Addition of the two *Veillonella*-conditioned media did not affect LDH (viability of the cells) or induce the secretion of the inflammatory cytokines CCL20, CXCL10, or IL-1β (under detection level of 15.6 pg/mL of ELISA assay for each respective cytokine). IL10$^{-/-}$ mice colonized with *V. parvula* resulted in a significant body weight loss, increased colon weight and colon mKC (IL-8 homologue) levels, and a reduction in caecum weight, indicative of intestinal inflammation (Fig. 3D). In agreement with the in vitro findings, *V. dispar* did not promote inflammation in IL10$^{-/-}$ mice (Fig. 3E–G). In contrast to previous studies with the pathobiont AIEC (ref. 35), no exacerbation of inflammation was observed in *V. parvula* pre-colonized IL10$^{-/-}$ mice subjected to 5 days with piroxicam as shown by distal colon weight (37.5 ± 7 vs 37.7 ± 11.1 mg/cm in piroxicam vs piroxicam colonized with *V. parvula*, respectively; $p > 0.05$; $n = 8$–10/group).

Host-related omics show only a moderate relationship with future relapse

Sampling mucosal biopsies allowed for simultaneous multi-omics of microbial and human cells from the same sites, which potentially can reveal host–microbe interactions. Host transcriptomics data from the same biopsies was first examined using principal component analysis (PCA), which highlighted modest, albeit significant, shifts in gene expression along the 6th and 9th PCs associated with future relapse (pooled samples, Wilcoxon test $p < 0.05$, Supplementary Fig. 3A). While differences did not remain significant along PC6 once we split samples by biopsy site (Supplementary Fig. 3B), we observed a significant difference for R samples along PC9 (Wilcoxon test $p < 0.05$). The only other variable associated with both PC6 and PC9 was patient gender(Wilcoxon $p < 0.05$; Supplementary Fig. 4). While PC1 values were not significantly associated with future relapse, they were strongly associated with sampling site, with TI samples being the furthest away from all other sites (Supplementary Fig. 3C, D, Wilcoxon test $p < 0.05$). There were also significant differences between TI and R samples along PC2, between R and AC samples along PC3, and between all biopsy locations along PC4 (Supplementary Fig. 3D).

Differential expression analysis was conducted on 21,477 genes, accounting for gender in the DESeq design. Out of the 21,477 analysed genes, only *MTND1P23* was differentially expressed when considering all samples together, with this gene found to be upregulated in patients who relapsed (LFC = 1.45, $q < 0.05$). Once we compared samples from individual biopsy sites, more genes were identified as differentially expressed, presumably due to the removal of intra-site differences from the signal. Here, 35 differentially expressed genes (DEGs) were identified between relapse and remission in TI samples, 10 in AC samples, 7 in DC samples (Supplementary Fig. 3E), and 1 in R samples (*CYP3A4;* $p < 0.05$).

PCA of qq-normalized beta methylation values showed significant shifts along PC3, associated with patient status (pooled samples, Wilcoxon test $p < 0.05$, Supplementary Fig. 5A; AC and TI samples, and Supplementary Fig. 5B), age, BMI ($p < 0.05$ for Spearman's correlation coefficient, Supplementary Fig. 5C), and sex (Wilcoxon test, Supplementary Fig. 5D). Both PC1 and PC3 were depicted in Supplementary Fig. 5E, F.

After filtering out failed, non-specific, and SNP-adjacent probes, we tested 709,163 CpG probes, and found 233 probes differentially methylated after FDR correction, whereof 34 were within promoter binding sites (Supplementary Data 4). As complex diseases often are associated with subtle methylation changes across regions (for example, promoters[36]), we used Gene Set Enrichment-based analysis to detect small differences that were consistent across the same region[37], resulting in 1561 genes differentially methylated at promoter binding sites (Supplementary Data 5).

As DNA methylation inherently controls gene expression, we investigated these associations and found that expression of 445 genes was significantly associated with methylation at promoter sites (Supplementary Data 6). Out of those 445 genes, 9 were differentially expressed between relapse and remission (pooled analysis, nominal p-value < 0.05) and differentially methylated at promoter binding sites (Supplementary Fig. 6A). Among these were HAVRC2 (a.k.a. TIM3[38]) and ICAM1[39] which have been implicated in UC pathogenesis, and GAPT[40] and KLHL6[41] known to be crucial in the life cycle of B-cells (Supplementary Fig. 6B).

Future relapse contributes to microbiome variance

As UC is a complex, heterogeneous, and multi-factorial disease, we next investigated how collected clinical variables (sex, age, antibiotics usage, and remission/relapse at diagnosis and at 6 months) contribute to explaining variance in omics data. We used distance-based redundancy analysis (dbRDA) followed by PCA[42]. In the PCA (Fig. 4A), the axes are constrained by linear combinations of exploratory variables and totally account for 16% of microbiota variance across 8 variables: PUCAI at diagnosis (2.6%), sex, BMI and relapse status (1.1% each), age, feeding mode, birth mode and antibiotic intake (<1%). The constrained axes for transcriptomics accounted for 43% of variance (Fig. 4B), with

https://doi.org/10.1038/s41467-025-62533-z

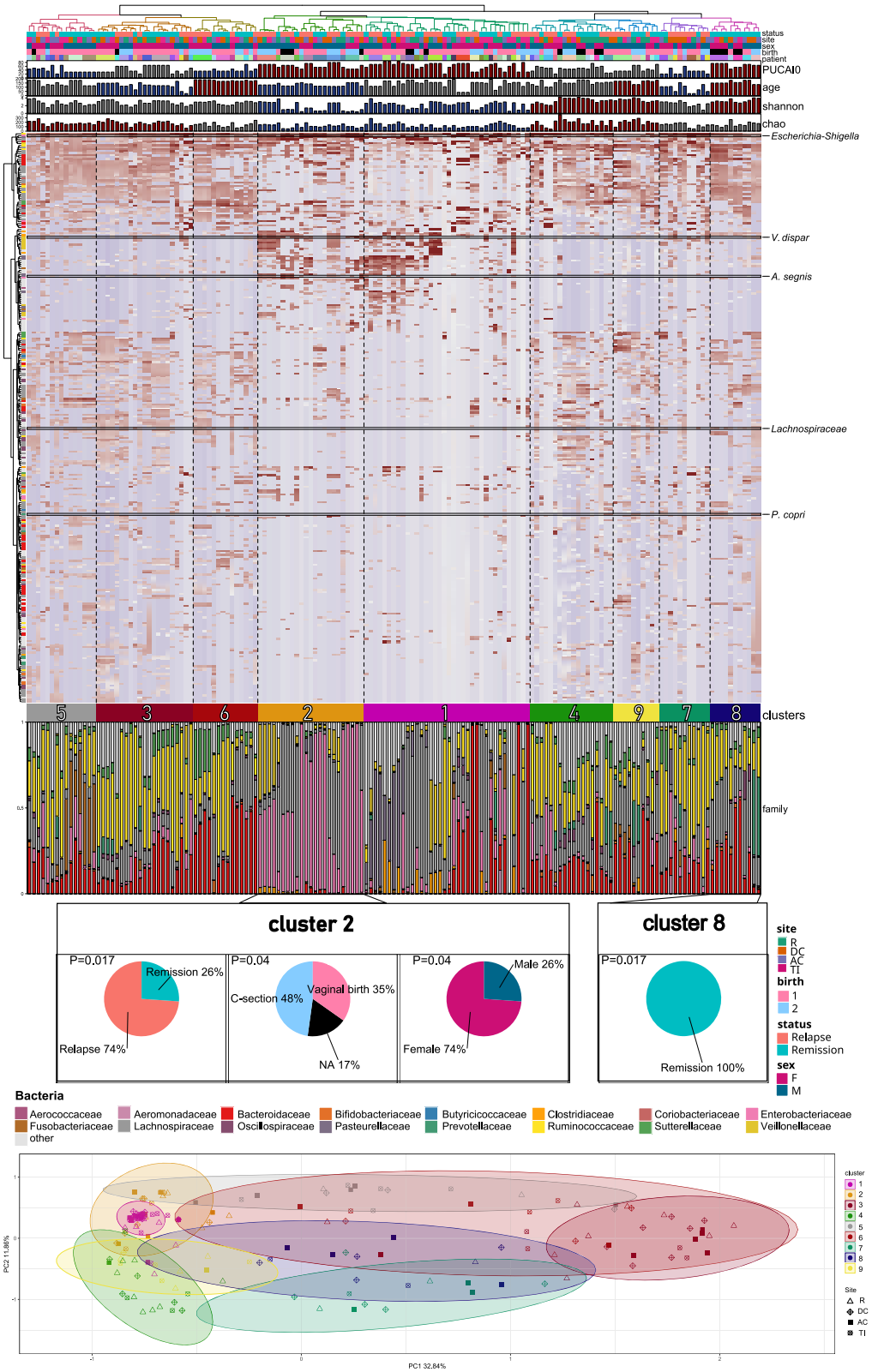

Fig. 2 | Heatmap of ASV abundance, normalized by patient. Clusters, presented in the PCA below, were identified with the dynamicTreeCut R package. Clusters with PUCAI at baseline, Chao1, Shannon, and age values significantly higher than the cohort average are indicated with red bars, while those with significantly lower values have blue bars. Indicator ASVs for clusters 2 and 8 (associated with relapse and remission at 6 months) are marked with lines on the heatmap. Bar plots underneath the heatmap indicate the relative abundance at the family level. All tests were two-sided, with no multiple testing correction.

https://doi.org/10.1038/s41467-025-62533-z

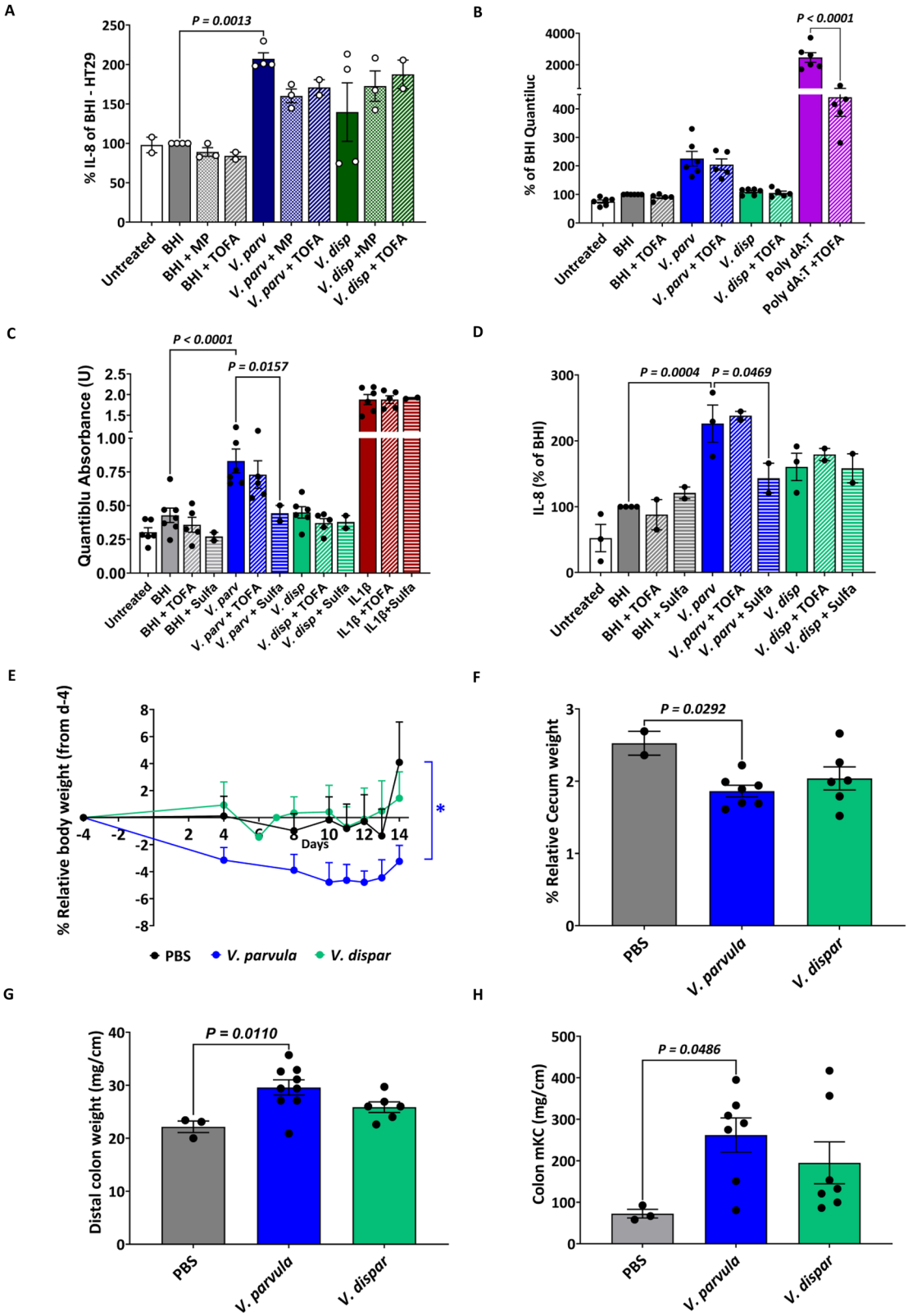

Fig. 3 | *V. parvula* induces pro-inflammatory response in epithelial cells and in IL10KO mice. A HT29 and **B**–**D** HCT116 Dual reporter cells cocultured with conditioned media of *V. parvula* (V. parv) and *V. dispar* (V. disp) and BHI-growth medium in the presence of IBD-drugs Tofacitinib (TOFA), methylprednisolone (MP), and Sulfasalazine (Sulfa), showing IL-8 secretion (**A**, **D**) and activation of NFkB (**C**) and IRF (**D**) pathways. Each dot represents an independent experiment. **E**–**H** IL10-/- mice were colonized with *V. parvula* and *V. dispar*, and body weight, caecum, and colon weight were measured, and levels of colon mKC. Each dot represents one IL10-/- mouse. ANOVA (one-way, two-sided) $^{*}p < 0.05$ followed by post-hoc analysis. Numbers in groups: **A** ($n = 2$–4 individual experiments), **B** ($n = 2$-6 individual experiments), **C** ($n = 2$–7 individual experiments), **D** ($n = 2$-4 individual experiments), **E** ($n = 3$–7 mice/group), **F** ($n = 2$–7 mice/group), **G** ($n = 3$–9 mice/group), and **H** ($n = 3$–7 mice/group). Error bands—95% mean confidence interval.

https://doi.org/10.1038/s41467-025-62533-z

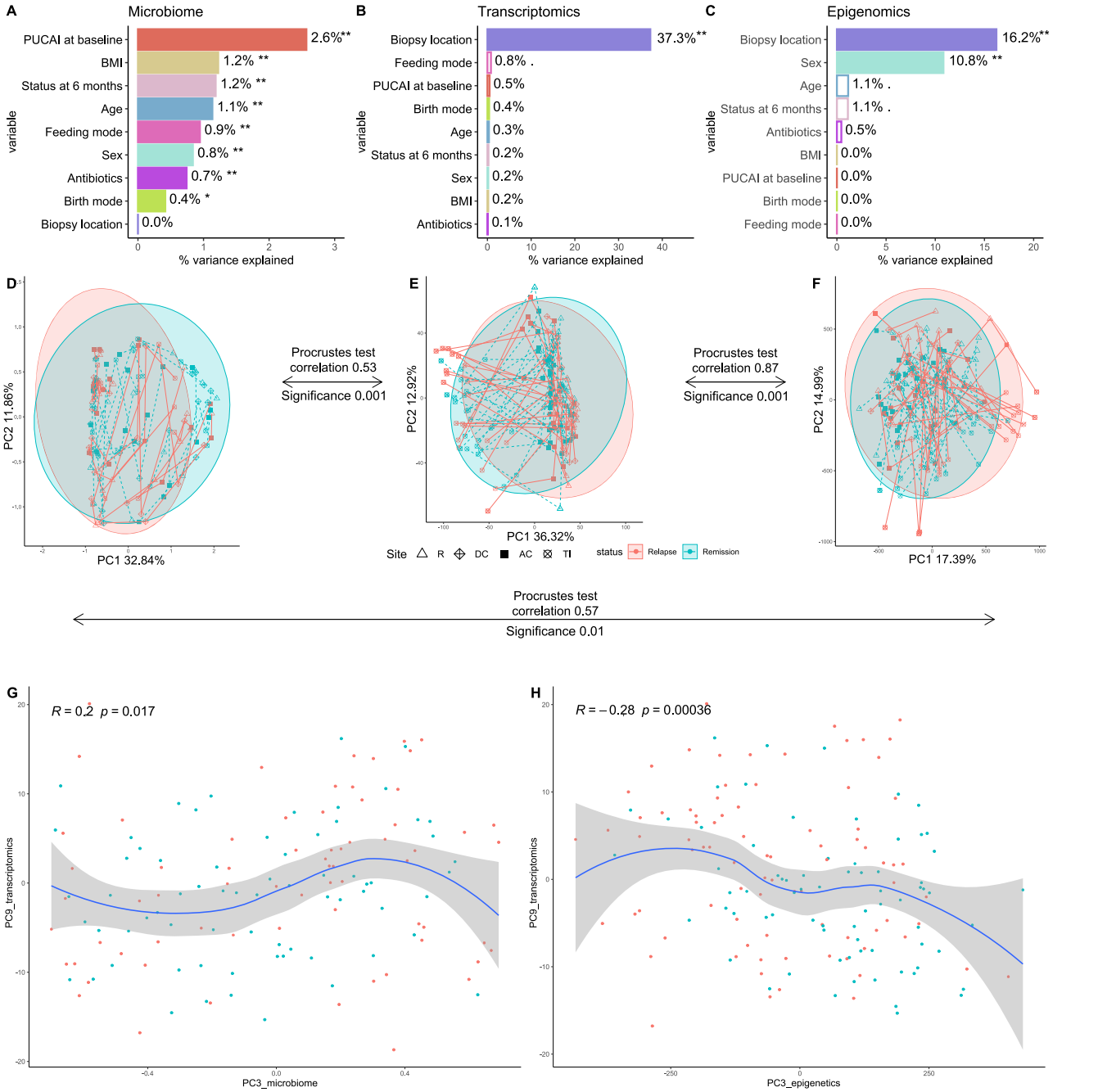

Fig. 4 | Multi-omic summary. A–C Percentage of variance explained in microbial, transcriptomic, and epigenomic data, respectively. **D–F** PCA plots of the first and second principal components for microbial, transcriptomic, and epigenomic data, respectively. The significance and degree of association between each of the omics are indicated in the PCAs. **G, H** Principal components associated with relapse and remission and with each other for microbiome and transcriptome (**G**) and transcriptome and epigenome (**H**). Significance codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1. All tests were two-sided, with no multiple testing correction.

biopsy site contribution of 37% and remaining variables for less than 1%. The epigenomics-constrained axes also accounted for 37% variance, and biopsy site, age, and relapse status at 1% each (adjusted $p < 0.1$, Fig. 4C).

Procrustes analysis of pairwise omics combinations showed significant associations between transcriptomics and the two other omics types (PERMANOVA $p = 0.001$), and between epigenome and microbiome ($p = 0.01$, Fig. 4D, F). Principal components associated with relapse (above) were significantly correlated with each other for microbiome and transcriptome (PC3 and PC9, respectively), and transcriptome and epigenome (PC3 and PC9, respectively, Fig. 4G, H).

The interferon gamma signalling pathway is overrepresented in the rectal transcriptome that is associated with relapse-linked bacteria

To identify specific host–microbiota interactions that may contribute to relapse, we conducted LASSO regression analysis to establish sets of transcripts associated with bacteria that were differentially abundant between future relapse and remission groups. For the remission group, 88 significantly overrepresented pathways were identified, containing 173 host genes in total, which were significantly associated with 7 bacterial ASVs. For the relapse group, 171 significantly overrepresented pathways were

present, containing 201 genes significantly associated with 11 ASVs (Supplementary Data 7). Of these, 141 and 58 pathways were unique to the relapse and remission group, respectively, whereof 33 selected unique pathways were associated with more than 1 ASV (Supplementary Data 7, see Methods section for details). The most prevalent functional (Reactome) assignments were either Immune System (Innate, Adaptive, or Cytokine Signalling) or infection response (15 pathways in total, Supplementary Fig. 7A). In particular, 4 ASVs were associated with interferon gamma signalling in the relapse group, whereof 3 ASVs were elevated in the relapse group: *A. rava*, *V. parvula*, and *B. thetaiotaomicron* (Supplementary Fig. 7A). In the relapse group, the potentially orally originating ASVs *A. rava* and *V. parvula* were associated with 68 and 53 transcriptomes, respectively. Specifically, these included GBP1, an interferon-induced GTPase involved in response to pathogens[43]; WWP1, an important regulator in TLR-mediated inflammation[44]; CXCL9, an interferon-induced chemokine involved in regulation of immune cell migration[45], and PROK2, a promoter of antimicrobial response[46]. In *V. parvula*, we also observed associations with BCL6, CEBPB, SOCS3, MLKL, TSC22D3, indicative of sustained type I/II interferon responses, innate immune transcriptional activation, and epithelial barrier dysregulation. *A. rava*, on the other hand, was associated with a distinct Th1-skewed transcriptional program (IL12RB2, STAT1, and GBP4) and oxidative and neuroimmune genes such as SOD2, GABRR2, and GRIN3A. Finally, the UC-implicated gene ICAM1, which we found to be under differential epigenetic regulation between relapse and remission groups (see previous section), was also associated with four ASVs (*A. rava*, *V. parvula*, *B. thetaiotaomicron*, and *P. excrementihominis*), with the first three ASVs being relapse-related (Supplementary Data 8 and Fig. 5).

All associations related to the immune system and infection response were present only in the relapse group, with the exception of *R. lactatiformans* and *F. prausnitzii*, associated with transcripts from the Interferon alpha/beta signalling pathway present for the remission group. On the other hand, associations unique to the remission group were mostly connected to normal metabolic functions.

Given the large contribution of sampling site to host transcriptome variation, subsequent GSEA[47] functional analysis was conducted on pooled results, as well as separately for each biopsy type. While none of the 1463 Reactome pathways were significantly enriched for DC differential expression (corrected *p*-value at 0.05), two pathways were significantly enriched for AC (Platelet sensitization by LDL and Response to metal ions), three for TI (Cytoprotection by HMOX1, Haem signalling and Response to metal ions), and one was significant for R (Aflatoxin activation and detoxification), and four narrowly missed statistical significance, including two interferon signalling pathways—gamma and alpha/beta (corrected *p*-value at 0.059, Supplementary Fig. 7B). For interferon signalling, 6 transcripts associated with bacteria abundance (ICAM1, GBP1, XAF1, GBP4, HLA-C, and STAT1) were also part of the core enrichment. Finally, pooled samples analysis revealed one significantly enriched pathway (response to metal ions).

Multi-omics integration with ML shows prognostic potential of host epigenetic and microbiome data

While single omics analysis showed associations with relapse, we further explored their combined relapse-predictive power by integrating

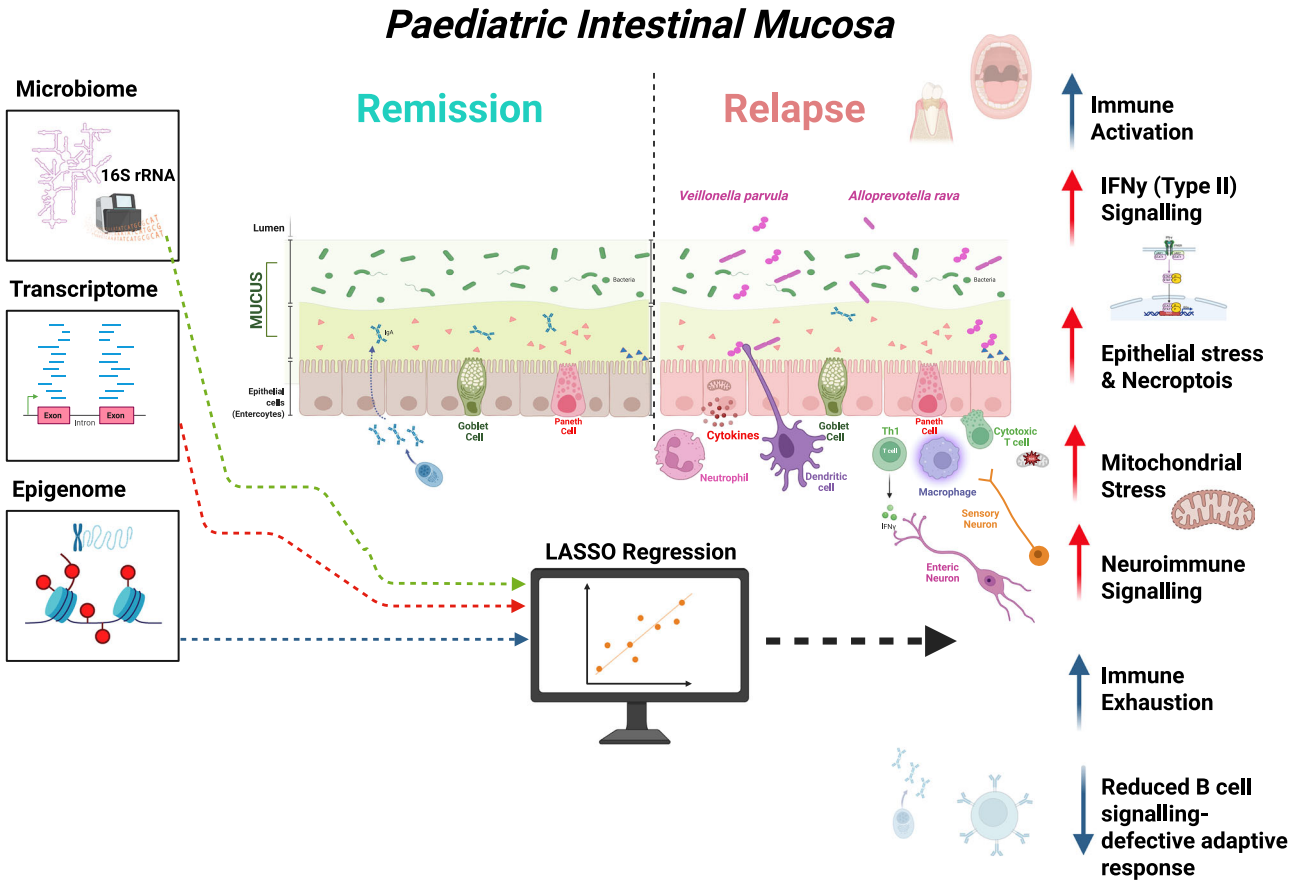

Fig. 5 | Overview of disrupted host–microbiome interactions in relapsing paediatric patients. The interactions were inferred from correlations between microbial abundances and host transcript expression using LASSO regression models. Created in BioRender. Mac Sharry https://BioRender.com/bki9rom.

available datasets using ML by employing the powerful ML algorithm eXtreme Gradient Boosting (XGBoost), which has already shown promise in many areas of omics research[14,48–50]. We also assessed the inclusion of some patient characteristics in the models, specifically patient age, mode of feeding (breast or bottle), and mode of delivery (vaginal or C-section).

To avoid issues with missing data, we initially only trained ML models on the 100-biopsy intersection for which all data types were available (32R, 30 AC, 30 TI, 8 DC). Models were trained on pooled samples from all biopsy sites, as well as on data from individual biopsy sites, except DC, which had too few samples. Models were subjected to a nested cross-validation (CV) performance assessment routine where an ensemble of 10 XGBoost models was used to make predictions (see "Methods").

As a baseline model, we also initially trained an ensemble of XGBoost models using only a patient's PUCAI score from baseline. We found that for each scenario considered, this model achieved either close to or worse than random performance (AUC range = 0.27–0.48, Supplementary Data 9 and 10). This, however, is not unexpected as we already showed no significant difference in baseline PUCAI between relapse and remission groups.

When using our multi-omics approach, we observed strong relapse-predictive capability from a model trained using both microbiome and epigenome features from pooled biopsy sites (AUC = 0.77, 95% CI: 0.68–0.86, Fig. 6A, B and Supplementary Data 9). Comparable performance was also seen when data was split by biopsy site (Fig. 6C and Supplementary Data 10), with a model trained on host transcriptome features from TI samples achieving the highest AUC (AUC = 0.80; 95% CI: 0.63–0.96; Fig. 6E). Additionally, good predictive performance was observed for models trained on AC (microbiome + patient characteristics; AUC = 0.74; 95% CI: 0.55–0.92) and R samples (epigenome + patient characteristics; AUC = 0.71; 95% CI: 0.53–0.89; Fig. 6C) when predicting relapse in these samples.

Given the complexity of these datasets, we wanted to further assess the stability and generalizability of our models by comparing the test set performance with that of the validation set, which is generally considered a good indicator of how a model will perform on potential external datasets. If a large gap exists between test and validation, it could indicate model instability. We estimated the validation performance by using predictions from the inner CV loop for those models outlined in Fig. 6. We observed only moderate differences between test and validation performance for models trained on pooled samples from all biopsy sites (Supplementary Fig. 8A). The highest test AUC achieved in this case was 0.77 and this model had a validation AUC of 0.88 (difference = 0.11). For these models, we found that the model with the highest test performance was also among those with the top validation performance and among those with the smallest difference between metrics, highlighting its potential stability. For individual site models, we again often saw the smallest differences between test and validation performances in those models with the top test AUC (Supplementary Fig. 8B). However, for some individual site models, we observed larger differences in performance between test and validation sets. For example, a model trained on epigenome features from TI samples and patient characteristics had a test AUC of 0.73 and a validation set AUC of 0.96 (Difference = 0.23; Supplementary Fig. 8B). Similarly, an R site model trained on epigenome and patient features, and an AC site model trained on host genotype, epigenome and patient features, had differences in performance of 0.3 and 0.27, respectively (Supplementary Fig. 8B). This could indicate that these models may not generalize as well to unseen data and may be less stable than other models highlighted in our analysis.

We also conducted an error analysis to see if potential model errors were linked to specific patient features. For the pooled site model with the highest performance (AUC = 0.77), we observed that more remission patients were correctly classified than relapse patients (Fisher's exact test $p = 0.008$; Supplementary Fig. 8C). This model also had a significantly different misclassification rate for mild (PUCAI score: 10–34) or moderate (PUCAI score: 35–64) UC at baseline compared to severe (PUCAI ≥ 65, classified according to[51]) UC and a different rate of misclassification based on mode of feeding (Fisher's exact test $p < 0.05$; Supplementary Fig. 8C). These differences could indicate that our model may be influenced by these patient-related features and highlights potential limitations in terms of the model's generalizability. The error analysis of models based on individual biopsy sites did not show any such differences in misclassification rates based on the available metadata features.

Up to this point, models had been trained on samples for which all data types were available in order to identify those data combinations that best predict relapse, using a consistent underlying sample set. However, as many of the top-performing models used only features from some and not all data types, additional samples were available on which we could further assess model performance. For this part of the analysis, we focused on the top three performing data combinations for each scenario previously considered (pooled, R, AC, and TI). The number of additional samples available for these data combinations are presented in Supplementary Data 11.

In this part of the analysis, we repeated our nested CV performance assessment routine using all available samples for the top three data combinations, to gain additional insight into their potential performance and stability (Supplementary Fig. 9 and Supplementary Data 11). This analysis was conducted for both the pooled sample and individual biopsy site models. For most pooled sample models, we observed little to no difference in performance between those trained on the full dataset and those trained on the initial subset (Supplementary Fig. 9A). Similar trends were observed for the top three models trained on R biopsy site samples with only the model trained on microbiome and patient features showing larger differences in performance (Full AUC = 0.562 ($n = 42$); Original AUC = 0.68 ($n = 32$); Supplementary Fig. 9B). However, for models trained on AC and TI samples, lower performance was generally observed for models trained on the full dataset compared to the initial training set (Supplementary Fig. 9C, D). In the case of AC models, we observed a reduction in AUC ranging from 0.12 to 0.26 for the previously top 3 performing models (Supplementary Fig. 8C). Similarly, in the case of the previously top performing model trained on host transcriptomic features from TI samples, we observed a reduction in performance of 0.21 (Supplementary Fig. 9D). However, for models trained using microbiome and epigenome or epigenome and patient features from TI samples, we observed only a reduction in AUC of 0.07 and 0.05, respectively.

Overall, omics data like the host epigenome and gut microbiome were found to be most predictive of relapse, with the stability and performance of models often improved by pooling samples from all biopsy sites together. We also found that many models could be improved by the inclusion of patient characteristics such as age, mode of feeding, and mode of birth.

Discussion

With the advancement of high-throughput sequencing and culture-independent microbial methods, multi-omic approaches to biomarker discovery and therapeutic outcome prediction have received increasing attention. This approach seems particularly warranted for UC, as abnormal interactions between host and microbiome are postulated to considerably contribute to disease development[10]. Addressing this challenge in a treatment-naïve paediatric UC cohort has the added advantage of targeting more recent pathophysiology while avoiding confounders caused by years of medication. We have shown that this approach may offer the opportunity to tailor the initial treatment of children at diagnosis according to the likelihood of future relapse. Most multi-omic approaches have focused on outcome prediction

https://doi.org/10.1038/s41467-025-62533-z

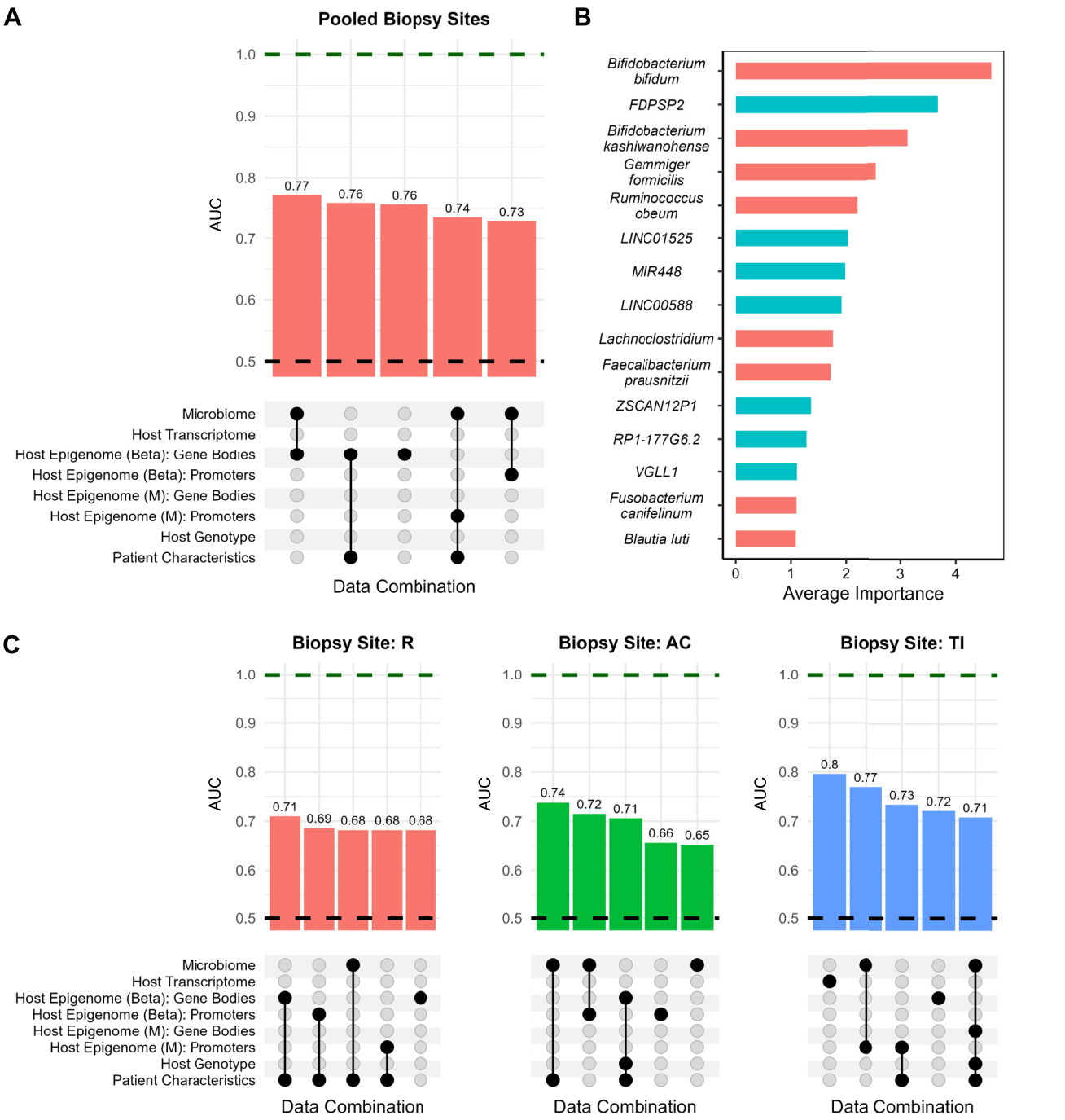

Fig. 6 | ML analysis to predict relapse 6 months after diagnosis. A Upset plot of the top five performing models based on the AUC metric when training an ensemble of XGBoost models on samples pooled from all biopsy locations. Dot-plots underneath highlight the data types used to train this model. Presented are only those models where the addition of a data type increased model performance. **B** Top 15 important features for the top-performing model in (**A**). Bars are coloured by the data type to which the feature belongs: 'Microbiome' (pink) and 'Host Epigenome (Beta): Gene Bodies' (blue). **C** Top five models for predicting relapse when samples were split by biopsy site. All AUC values shown in this figure are based on the outer loop of our nested CV framework, which was used to assess model performance. Patient characteristics: age in months, breast or bottle fed, and mode of birth (vaginal or C-section).

after biologic therapy (reviewed in ref. 52), until the PROTECT study[15], where *Ruminococcaceae* and *F. prausnitzii* from rectal biopsies were associated with corticosteroid-free remission at 52 weeks in treatment-naïve paediatric patients with UC. In our study, based on absolute microbial abundances, we similarly found *R. obeum, R. torques*, and *F. prausnitzii* underrepresented in the mucosal microbiome of patients relapsing within 6 months. In particular, *F. prausnitzii* has displayed anti-inflammatory properties by upregulating anti-inflammatory cytokines[53], and we found both this species[54] and *R. obeum*[55] to be depleted in the microbiome of patients with UC. On the other hand, *R. torques* is a known mucin degrader previously implicated in the pathogenesis of IBD[56]. This indicates that while certain bacteria might contribute to UC development, their presence might also contribute to treatment susceptibility.

A major group of bacteria linked to future relapse in our study was orally associated and included *V. parvula, Fusobacterium* genus, and *A.*

rava. While clustering is inherently sensitive to distances and methods used, other oral bacteria, *A. segnis* and *V. dispar*, were also indicator species in the relapse-related cluster 2, further strengthening this observation. Ectopic gut colonization by oral microbiota has been implicated in various gastrointestinal diseases (reviewed in ref. 57). Normally, the resident gut species would prevent colonization, but in UC, bacterial diversity is depleted[13], opening new niches for potential colonizers from the upper gastrointestinal tract. In our study, the relapse group was characterized by lowered species richness, which has previously been reported for non-responders to steroid therapy[58]. While the exact mechanisms of oral bacteria's pathogenicity in the human gut remain largely unknown, many of them (e.g., *Veillonella* and *Aggregatibacter*) produce pro-inflammatory H_2S[57]. Inversely, paediatric patients who remained in corticosteroid-free remission after 52 weeks were depleted of oral cavity microbiota[15]. This suggests that reducing oral colonizers from the gut could potentially be key to successfully maintaining remission.

Out of the bacteria characteristic of the oral cavity, two *V. parvula* ASVs were overrepresented in the relapse group, while *V. dispar* was an indicator species for relapse-related bacteria cluster 2. Mice monocolonized with bacteria from this genus were characterized with repressed production of IL-22[59], a cytokine involved in immune regulation and inflammatory responses in UC[60]. In particular, *V. parvula* colonizes the gut through inflammation, by switching metabolism to nitrate respiration (utilizing the *narGHJI* operon)[61]. Schirmer and colleagues also found it to metabolize immuno-suppressants[62]. Hence, we further investigated two *Veillonella* species (*V. dispar*, *V. parvula*) for their translational properties, and found that *V. parvula* induced inflammation in $IL10^{-/-}$ mice and their conditioned media induced epithelial cell IL-8 secretion and NFκB activation. These findings are in line with previous studies on a moderate secretion of IL-8, IL-1β, IL-10, and TNF-α by dendritic cells co-cultured with *V. parvula*[63] and on *V. parvula*-LPS-induced inflammation in macrophages via TLR4 and in impairing colonic barrier function[64]. In contrast, *V. dispar* did not provoke major inflammation in epithelial cells or mice. While both *V. parvula* and *V. dispar* are common members of the oral microbiota, emerging evidence indicates that *V. parvula* may have a more pronounced role in mucosal inflammation, particularly in chronic airway disease[65] and IBD[61,64]. As observed in our cell line and murine models, the inflammatory nature of *V. parvula* is evident. This may be due to differences in cell membrane composition, adhesions, metabolic byproducts, or interactions with host pattern recognition receptors. Indeed, *V. parvula* and *V. dispar* inhabit differing niches in the oral cavity, gingival plaque, or saliva[66] and display distinct genomic differences. Further comparative studies are needed to delineate the species-specific contributions to disease pathogenesis. Inflammation induced by DSS (dextran sulfate sodium) was reported to promote the engraftment or expansion of *V. parvula*[61]. However, when using genetically modified mice, $IL10^{-/-}$, we were not able to validate this finding. This could be due to the stimuli used to promote inflammation in the two studies (NSAID vs DSS) or the low degree of basal inflammation in our $IL10^{-/-}$ colony. When we examined whether IBD drugs could reduce *Veillonella*-induced responses, only sulfasalazine was able to reduce IL-8 and NFκB activation on epithelial cells, while a steroid and a JAK inhibitor did not. These results extend other findings on *V. parvula* affecting thiopurine metabolization, indicating this bacterium may be relevant for treatment failure in patients with IBD.

In contrast to the gut microbiome, the host epigenome and transcriptome had a weaker relation with sustained remission. Only 233 CpG probes were found to be differentially methylated between relapse and remission groups. A much larger number of genes (1561) were differentially methylated at promoter sites. Patient status at 6 months did not significantly contribute to variance explanation in transcriptomics or epigenomics but was instead associated with lower-order principal components, which is in line with the PROTECT cohort, where only 33 differential genes were associated with relapse at 52 weeks[25]. Interestingly, three of these genes (*C4BPA, CA1*, and *AQP8*) were also differentially expressed between relapse and remission in TI samples in our study. Those genes, as well as another differentially expressed gene in TI from our study, *CA2*, are members of the corticosteroid response gene signature, identified as a corticosteroid response predictor in the PROTECT cohort and further validated in adult UC patients[67]. Others have identified co-expressed gene modules, associated with treatment outcomes, but provided no further characteristics[68]. Taken together, our transcriptomics results implied a weak signal associated with patient status and mainly related to genes associated with bacteria–host interaction. This was further corroborated by procrustes analysis, where microbiome composition and transcriptome were significantly correlated with each other, as well as PC correlation analysis, where PCs associated with relapse in microbiome and transcriptome were also significantly correlated.

Further analysis pinpointing specific host–microbiota interactions showed that transcripts for interferon gamma signalling were associated with the abundance of four bacterial ASVs, either up (3) or down (1) in relapse patients. Early mouse studies have suggested that interferon gamma presence is crucial to the development of IBD[69], and it has been speculated that overproduction of interferon gamma is a result of abnormal immune response to commensal bacteria, which leads to disruption of both epithelial[70,71] and vascular[72,73] barrier in the intestines. On the other hand, the exact impact of interferon gamma could also be dependent on the host's genetic makeup[74] and be less pronounced in UC, compared to Crohn's disease[75]. In our study, the relapsing patients' rectal transcriptome is also enriched in transcripts associated with interferon gamma signalling, suggesting at least a role in treatment response, even if not pathogenesis. One of the proposed mechanisms of glucocorticoid (GC) immunosuppressive action is through inhibition of Janus kinase-STAT signalling activation by interferon gamma[76]. Therefore, increased expression of transcripts from this signalling pathway (possibly driven by the presence of the aforementioned four bacterial ASVs in the gut) might lead to hampering of this process and higher the chances of relapse on steroid treatment. Another interferon-related signalling pathway, involving interferons alpha and beta, was enriched among transcripts significantly associated with *F. prausnitzii* and *R. lactatiformans* in the remission group. Variants of interferon beta were previously proposed as a treatment for UC, but the results were inconclusive[77,78]. Both interferons were found to be protective against colitis in animal studies[79,80] and were also crucial in the healing of colitis-induced mucosal damage[80]. *F. prausnitzii* and *R. lactatiformans* were overrepresented in patients within the remission group, and it is therefore plausible that they contribute to sustained remission via interferon alpha and beta modulation.

V. parvula and *A. rava*, typically considered members of the oral microbiota, were detected at increased abundance in inflamed intestinal biopsy specimens from relapsing patients, supporting prior studies that have highlighted microbial translocation or ectopic colonisation as hallmarks of IBD-associated dysbiosis. Importantly, their presence correlated with divergent host transcriptional responses, with shared activation of interferon-regulated genes and epithelial stress markers. *V. parvula* was associated with expression of BCL6, CEBPB, SOCS3, GBP1, ICAM1, MLKL, TSC22D3, and WWP1, indicative of sustained type I/II interferon responses, innate immune transcriptional activation, and epithelial barrier dysregulation. The elevated expression of SOCS3[81] and TSC22D3[82] further suggests that compensatory anti-inflammatory pathways are engaged but may be insufficient to resolve inflammation in the context of persistent microbial and immune stimulation. Notably, MLKL upregulation points to activation of necroptosis, a programmed form of inflammatory cell death that has been implicated in epithelial damage and disease relapse in IBD (reviewed in ref. 83). Conversely, the presence of *A. rava* was

associated with a distinct Th1-skewed transcriptional program, including IL12RB2, STAT1, GBP1, GBP4, CXCL9, and oxidative and neuroimmune genes such as SOD2, PROK2, GABRR2, and GRIN3A. These findings align with enhanced IFN-γ signalling, increased T cell chemotaxis[84], and mitochondrial stress[67,85]—features previously associated with disease severity and relapse risk in paediatric IBD. The involvement of neuronal signalling genes also raises the possibility that *A. rava* may contribute to neuromodulatory processes such as visceral hypersensitivity and gut–brain axis dysfunction. Despite their taxonomic and functional differences, both microbial taxa appear to converge on a shared interferon-γ–driven immune axis, with GBP1 regulation serving as a key node of overlap. These results support a model in which microbial dysbiosis in paediatric IBD relapse fosters chronic mucosal immune activation, stress signalling, and impaired epithelial repair. Further, they highlight the potential for microbiota-derived immune signatures to serve as biomarkers of relapse-prone disease states and as mechanistic targets for therapeutic intervention.

To our knowledge, most studies of relapse in paediatric UC only gathered clinical or biochemical data at baseline, which proved to possess little to no predictive power[21]. These included severity measured by PUCAI at baseline, which, as in our study, had no predictive value, necessitating investigation of alternative measures. The PROTECT study did use multi-omic features, but only to a limited extent, whereby antimicrobial peptide signatures and log-transformed relative abundances of *Ruminococcaceae* and *Sutterella* were added to the clinical model (haemoglobin concentrations and remission at week 4), improving AUC from 0.68 to 0.75[25]. A follow-up study showed that including *IL13RA2* gene expression in the model for the same cohort could eliminate the need to observe remission at 4 weeks[86], achieving an AUC of 0.7.

Our study is unique in that we assessed the predictive ability of all possible combinations of four omics datasets generated from biopsies across four intestinal locations. Furthermore, we analysed the absolute, instead of the relative microbial abundance, which makes validating potential findings more relevant and generalisable. We observed the best performance when combining microbiome and host epigenome data from pooled sites, with many models improved by the inclusion of patient characteristics. This overall performance was based on baseline sampling only and still achieved higher AUCs than previous studies. *Ruminococcus obeum* was among the most important features for models incorporating microbiome data, along with previously UC-associated *F. prausnitzii*[15,54], *Bifidobacterium bifidum*[87,88], and *Gemmiger formicilis*[14]. These results are in contrast to those obtained in the adult UC cohort[89], where no differences at baseline were found between responders and non-responders to corticosteroid treatment. On the other hand, the authors have spotted some trends similar to ours, such as "oralisation" of gut microbiome for non-responders. Intriguingly, oral bacteria such as *V. parvula* were also found to be associated with poor response to biologics therapy in adults[90,91].

Many of our top epigenome-based models also incorporated methylation of gene body regions, whereas most DNA methylation studies target the promoter regions due to the well-established effect on gene expression[92]. As a result, gene body methylation is often overlooked despite its significant, albeit less understood, role in the regulation of gene expression[92–94]. Among the most important gene body regions in our models were *FDPSP2, LINC01525*, miRNA-448, and *LINC00588*, which slightly improved performance over the PROTECT study. We further expanded on the traditional ML analysis by evaluating the model stability and conducting crucial error analysis. As a result, we identified several subgroups of patients for whom predicting future relapse status was more challenging. Future studies should focus on such groups of patients so that we can improve the accuracy of these relapse predictions.

Although our ML approach shows promise in terms of predicting future relapse status, there were some limitations to our study. While we had a rather large number of samples in our study, not all omics types were available for all samples, requiring intersectional subsampling. We therefore applied a nested CV methodology to make the most of the traditional training, validation, and test splits used in ML, as unfortunately, no suitable external dataset currently exists to validate our findings. Additionally, to maximize our sample size, many of our top-performing models were trained on samples pooled from all biopsy locations, thus, multiple samples from the same patient were included in the training data. As a result, while these models were trained on a rather large number of samples, the number of subjects included in the training data remains limited. While the PROTECT study[25] had a similar cohort and data types, it assessed relapse at 52 weeks after diagnosis, too far off from our 26-week end-point. Also, the 16S rRNA amplicons used a different primer set and sequencing approach, which was not weighted to absolute abundances. While our findings will hopefully be validated on new subjects with closer methodological alignment in the future, we were still able to conduct several ML checkpoints during our analysis (test vs validation performance, error analysis, model testing on additional samples from our cohort). This allowed us to assess the stability and robustness of our models while also highlighting some potential limitations of our models, something which is not always done in this type of ML analysis.

To conclude, we have demonstrated that a multi-omic ML model has the potential to predict patient outcome, despite comparatively weak signals from single-omic datasets. We also found an association between the pro-inflammatory oral bacteria *V. parvula* found in the ileo-colonic mucosa and treatment failure in paediatric UC, which warrants further investigations in multi-centre studies. Given the involvement of *V. parvula* and *A. rava* in both intestinal and oral environments, the oral microbiota may represent a practical, non-invasive window into gut immune status in paediatric IBD. Longitudinal assessment of oral mucosal inflammation and dysbiosis using oral swab-based 16S rRNA/Shotgun sequencing, qPCR panels, or salivary cytokine profiling could complement stool or biopsy-derived microbial data.

Methods

All research in this article was conducted following relevant ethical guidelines. The London - City Road & Hampstead Research Ethics Committee granted approval (reference 15/LO/2127) to collect research biopsies, including genetic material, at Barts Health NHS Trust, Whitechapel, London. The protocol was approved by the Digestive Diseases Bioresource Committee at Queen Mary University of London and Barts Health NHS Trust. Written informed consent was provided by all patients and their parents.

All animal studies were approved by the Animal Experimentation Ethics Committee (AEEC; application no. #2023/008), University College Cork, and by the Health Products Regulatory Authority (HPRA, project no. AE19130/P196).

Patient characteristics, sample storage, and inclusion criteria

Patients were children aged 2–16 years who were admitted for colonoscopy to diagnose UC. Children suspected of having UC under 2 years were not included to reduce the possibility of recruiting children with monogenic inflammatory bowel disease (IBD), who were later not found to have UC. Parents gave written consent for additional biopsies to be collected during endoscopy. Biopsies at diagnosis were immediately immersed in cryovials containing 1 mL RNA*later*™; they were stored at −70 °C in the Digestive Diseases Bioresource, which is a tissue bank registered with the UK Human Tissue Authority (HTA) under Barts Health NHS Trust and Queen Mary University of London. Research ethics committee approval (reference 15/LO/2127) had been granted to collect research biopsies, including genetic material. Frozen samples were transferred to University College Cork using approved carriers in accordance with HTA regulations. Children were managed

according to the ESPGHAN-ECCO guidelines[19,95], and no changes were made in their management because they were part of the study, other than taking consent for research samples at endoscopy.

As is normally seen in paediatric practice, colitis was extensive and severe in all examined patients, who were given steroids. All patients had pancolitis., whereof six patients also had inflammation extending to TI. No further biopsies were taken after treatment, as it is not a standard practice to perform a second endoscopy in children to directly assess mucosal healing[95].

Children were treated according to ECCO/ESPGHAN guidelines with aminosalicylates and corticosteroids to achieve remission. Children on corticosteroids were prescribed omeprazole concurrently to avoid gastritis. Some children with mild disease, did not require corticosteroids and were successfully treated with aminosalicylates only. Children who did not receive corticosteroids were excluded from the study. Children who failed to enter remission on corticosteroids were also excluded. Corticosteroids were tapered to zero within 3 months according to the guidelines[95]. After corticosteroid therapy commencement, 73% children achieved initial remission within 1 month.

All children who entered the study, therefore, achieved remission within 6 months. Children were deemed to have remained in remission if, after 6 months, they required no further corticosteroids (which equates to a paediatric UC score (PUCAI) of less than 10. Children were deemed to have relapsed if they required a new course of corticosteroids (with a PUCAI of greater than 10). The median relapse time was at 2 months (maximum−5 months). The primary endpoint of the study was, therefore, whether they were either in remission or had relapsed by 6 months after starting corticosteroids for the first time.

Of the 74 patients recruited into the study, 9 were excluded due to a lack of confirmed UC diagnosis. Furthermore, 7 patients were excluded due to a lack of corticosteroid treatment, and 2 patients were excluded for not achieving remission after initial corticosteroid treatment. Microbiome analysis was conducted for 48 patients (159 biopsy samples), epigenome for 55 (168 samples), and transcriptome for 53 (173 samples). The total number of patients in all analyses was 56. Further cohort characteristics are presented in Table 1 in the Results section. In the full cohort, there were no statistically significant differences between the relapse and remission groups (Welch's two-sample *t*-test and Fisher's exact test for quantitative and qualitative variables, respectively).

Sample processing, sequencing, and array preparation

Biopsy processing and nucleic acid extraction followed protocols described in Ryan et al.[14]. Briefly, biopsies in RNA-later were completely defrosted before performing DNA/RNA purification using the AllPrep DNA/RNA Mini kit (Qiagen, Valencia, CA, USA). They were then extracted from the RNA-later and moved into a tube with 350 μL of RLT buffer with β-mercaptoethanol (Sigma-Aldrich, St Louis, MO, USA), three 3.5 mm glass beads, and 0.25 mL of 0.1 mm glass beads (Biospec, Bartlesville, OK, USA). Disruption and homogenization were performed in a MagNA Lyser (Roche, Penzberg, Germany) twice for 15S at 6500 rpm, followed by DNA/RNA purification following the kit manufacturer's instructions. Purified genomic DNA was finally eluted in 100 μL of EB buffer, while RNA was eluted in 60 μL of RNase-free water.

Library preparation of 16S hypervariable region V3-V4 amplicons was also performed according to methods described in Ryan et al.[14], whereby PCR amplification of the V3-V4 hypervariable region of the 16S rRNA gene was conducted for aliquots of 200 ng of extracted DNA in a total volume of 30 μL. The primers (forward TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGCCTACGGGNGGCWGCAG; reverse GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGGACTACHVGGGTATCTAATCC) had Illumina adaptors with overhang nucleotide sequences added to the gene-specific sequences[96] and were used at a concentration of 0.2 μM. PCR amplification with the Phusion High-Fidelity DNA polymerase (Thermo Scientific, Waltham, MA, US) was conducted on a 2720 Thermal Cycler (Applied Biosystems, Foster City, CA, USA) under the following conditions: 98 °C for 30 s, proceeded by 30 cycles of 98 °C for 10 s, 55 °C for 15 s, 72 °C for 20 s and a final cycle of 72 °C for 5 min. The presence of the amplified 16S rRNA gene band was verified in agarose gels. Post-PCR products were purified using Agencourt AMPure XP magnetic beads (Beckman-Coulter, Brea, CA, USA) and eluted in 50 μL of EB Buffer (Qiagen, Venlo, The Netherlands). After purification, 5 μL of DNA was amplified in a second PCR using Nextera XT Index primer (Illumina, San Diego, CA, USA). This PCR was run at 98 °C for 30 s, followed by 8 cycles of 98 °C for 10 s, 55 °C for 15 s, 72 °C for 20 s, and a final cycle of 72 °C for 5 min. A second purification step with Agencourt AMPure XP magnetic beads was carried out after the Nextera PCR. The 16S V3-V4 rRNA gene amplicons containing the Nextera indexes were finally eluted in 25 μL of EB Buffer, and DNA concentrations were measured using the Qubit HS dsDNA assay kit (Thermo Scientific). Pooled libraries were created by adding 50 ng of each sample. Finally, diluted samples of the libraries were sent for sequencing at Teagasc Food Research Centre on an Illumina MiSeq for 2 × 300 bp reads. DNA extractions were used for qPCR assays as well, which consisted of a 16S rDNA qPCR in a SYBR-green assay, using a LightCycler480 (Roche) with 16S primers reactions and chemistry[97], normalized by 16S rDNA qPCR bacterial load estimates based on a calibration curve of 10-fold diluted *S. aureus* cultures as described in Patel et al.

Host RNA library preparation and sequencing were conducted as previously described[14], where RNA samples were first treated with DNase to remove any DNA traces employing the Turbo DNA-free kit (Ambion, Austin, TX, US) following the manufacturer's instructions. To determine RNA integrity, 1 μL of each RNA sample, obtained after DNase treatment, was loaded on RNA 6000 Nano Chip (Agilent Technologies, Santa Clara, CA, USA) according to the manufacturer's protocol. RNA integrity number (RIN), rRNA ratio, and concentration were measured by microfluidic capillary electrophoresis in a 2100 Bioanalyzer system (Agilent Technologies). Quality was considered acceptable if RIN was ≥6 and rRNA ratio ≥ 1.5. The sequencing was conducted with 2 providers: Edinburgh Genomics and Source Bioscience. At Edinburgh Genomics, samples were normalized to 1 ug of Total RNA in 50ul. Libraries were prepared using TruSeq stranded mRNA (Illumina) Enrichment and cDNA Synthesis on the Biomek FX (Beckman), with 10 PCR cycles. Post library clean-up was done on Biomek NX, and quality control of these cleaned libraries was done using the HSD1000 tapestation reagents (Agilent) with the samples diluted 1:5. Completed libraries were then normalized to 5 nM and quantified using Quant-iT dsDNA HS Assay Kit (Life Technologies, Carlsbad, CA, USA). Samples were pooled equimolar (based on Quant-iT plate reader values) and the resulting pool quantified by qPCR (KAPA Library Quantification Kit−Universal, Roche) before loading on NovaSeq at 100 paired-end reads. At Source Bioscience, the libraries were prepared using the Illumina TruSeq® Stranded mRNA Library Prep according to the manufacturer's protocol. During this process, the libraries were indexed using IDT for Illumina−TruSeq RNA UD Indexes. The prepared libraries were quantified via a fluorometric method involving a Promega QuantiFluor dsDNA assay; and qualified using electrophoretic separation on the Agilent TapeStation 4200. They were sequenced on the NovaSeq6000 platform using an S4 Flow-Cell with v1.5 chemistry to generate 150 paired-end reads.

The genotyping using microarrays was conducted according to the protocol as previously described[14]. Prior to genotyping, DNA aliquots were sent on dry ice to Macrogen (Seoul, South Korea) to be assayed on an Infinium ImmunoArray-24 v2 BeadChip. Intact genomic DNA was diluted to 50 ng/μL based on concentrations measured using Quant-iT Picogreen. All samples were hybridized on the microarray

according to the manufacturer's instructions. Briefly, whole genome amplification (×1000) was carried out on 200 ng of DNA. Subsequently, the products were fragmented, precipitated, and resuspended in a formamide-containing hybridization buffer. Next, samples were denatured at 95 °C for 20 min, and then placed in humidified containers for a minimum of 16 h at 48 °C, allowing SNP loci to hybridize to the 50mer capture probes. Following hybridization, the BeadChip/Te-Flow chamber assembly was placed on the temperature-controlled Tecan Flowthrough Rack, where subsequent washing, extension, and staining steps were performed. After staining, the slides were washed with a low salt wash buffer, immediately coated, and then imaged on the Illumina iScan Reader. Image intensities were extracted using Illumina's GenomeStudio software.

In contrast to the Illumina 450k[14], we here applied the larger Illumina MethylationEPIC array. The Illumina Infinium MethylationEPIC array BeadChip (850 K) was carried out by the Epigenomic Services from Diagenode (Cat no. G02090000, Seraing (Ougrée), Belgium). The DNA was deaminated with the EZ-96 DNA Methylation Kit (Zymo Research) according to Illumina's recommended deamination protocol. Bisulfite conversion was then controlled by qPCR. One assay targeting a methylated region of DNAJC15 and two assays targeting the GNAS locus (one assay for the unmethylated allele and one assay for the methylated allele) were used for quality control. Deaminated DNA derived from blood was amplified in parallel and served as a positive control. A sample passed the quality control when the received CT-value, either for the two GNAS loci or the DNAJC15 locus, reached the threshold not later than 5 cycles compared to the positive control.

Statistics and reproducibility

Our study followed an observational and cross-sectional study design (1 time-point). No statistical method was used to predetermine sample size, consecutive paediatric patients diagnosed with UC were recruited into the study instead. Patients were excluded from the study due to predefined exclusion criteria, described in the "Patient characteristics" section of the methods. No data were excluded due to low quality. As this was not an experimental study, the patients were not randomized. In mouse experiments, before the start of the experiment, mice were randomised to the different groups based on their age at the start of the experiment.

Microbiota sequencing and analysis

Quality of reads was examined using fastqc[98] version 0.11.8. Removal of Illumina sequencing adaptors was conducted with trimmomatic[99] version 0.39, leaving only sequences longer than 50 bases after trimming. Cutadapt[100] version 4.2 was used to remove primers. Subsequent analysis steps were conducted with DADA2[101] R package version 1.22. The sequences were right-side trimmed at 15 bases, with maximum expected errors set at 2 and 3 for forward and reverse sequences, respectively. The composition of the samples was inferred with the dada function, and long ASVs (amplicon sequence variants) were created by merging paired-end reads. Only ASVs between 390 and 460 bases long were kept for further analysis. Contaminant sequences were identified with decontam[102] R package, using blank samples and negative extraction controls. Chimeric sequences were removed with the removeBimeraDenovo function. ASV taxonomic classification up to genus level was conducted with mothur[103] version 1.44, using the SILVA database[104] version 138. SPINGO[105] version 1.3 was used for species-level classification. Species were assigned only for sequences classified at least to genus level by mothur. RDP database dispatched with SPINGO was used, unless only SILVA gave a non-ambiguous assignment. In case of genus-level discrepancies between databases, blast results from the NSBI 16S rRNA database were used. Counts, normalized by 16S rRNA gene copy numbers, were obtained with PICRUSt2[106]. Relative abundances were computed and used to better approximate absolute abundances from 16S qPCR results (CFUs/mL). Shannon and Chao1 diversity indices were computed with the iNEXT[10] R package version 3.0.0. Bray–Curtis distance measure was computed with the vegan package version 2.6[107].

Mann-Whitney U test (for pooled data) and mixed-effects models were used to determine differences in indices. Negative binomial zero-inflated mixed-effects model from NBZIMM[108] package version 1.0 was used to compute differential ASVs abundances. Status at six months was used as a fixed effect, while patient identifier was used as a random effect. dbRDA was used to determine the full percentage of variance explained by environmental variables, while the percentage of variance attributable to specific variables was determined by a procedure described in a tutorial for the VpThemAll package version 0.0.0.9000[109]. Variation partition tests were performed for each variable in the metadata, using the *test.varpart.wrap* function. *p*-values were adjusted for multiple testing, using the Benjamini–Hochberg prcedure[110]. Only the unique (corrected for other variables in the test set) portion of variance explained was taken into account for this procedure. Feeding mode was coded as a three-level variable for this purpose: breastfeeding, mixed breast and bottle feeding, and bottle feeding.

Patient microbial samples were clustered using Bray–Curtis distance and Ward's clustering criterion. Cluster assignments were performed with the cutreeDynamic[111] dynamic function, using default parameters except for minimum cluster size, which was set to 10. Differences between cluster and overall proportions for quantitative variables were tested with one-sample chi-squared tests. Differences between cluster and overall means for qualitative variables were tested with one-sample *t*-tests. In order to avoid bias, each patient's observation was counted only once per cluster, even if they had multiple samples within a cluster. Indicator families and ASVs for each cluster were identified with the multipatt function[112] using correlation-based indices and accounting for a differing number of samples within each cluster.

Epigenome analysis

Raw signal data was preprocessed by the preprocessRaw function from minfi[113] package version 1.44. Probes near SNPs were removed with the dropLociWithSnps function, while the getBeta function was used to calculate the beta values. Failed probes were identified with the detection function and subsequently removed. Non-specific probes were removed according to a previously described strategy[114], with previously generated datasets[115,116]. The BMIQ[117] procedure, implemented in watermelon[115] package version 2.4, was utilized to remove the bias from the type-2 probes. Batch effects were removed with ComBat[118] functionality from the sva package version 3.46, using the Sentrix ID variable as batch and the model matrix with patient status, gender, age, and sampling site. Quantile normalisation using the qqnorm function from base R was used before conducting principal components analysis with the prcomp function. Association of principal components with patient status was assessed with Mann–Whitney *U*-tests.

Differentially methylated sites were identified with linear mixed-effects models from lmerTest[119] R package version 3.1. Overall methylation across promoter sites of known genes was assessed with mCSEATest from mCSEA[37] package version 1.18, using the *t*-test statistic from a linear mixed-effects model as an enrichment measure (formula $\sim status + sex + age + Sentrix\_ID + Sentrix\_Position + Sample\_Group + (1|person)$). Correlations with expression values were computed with mCSEAIntegrate from the same package (only negative correlation results are reported).

Transcriptome analysis

Raw host RNA-Seq reads were first quality checked using FastQC v0.11.9[98] and MultiQC v1.14. Adaptors and low-quality reads were removed using Trimmomatic v0.39[99] with the following parameters:

SLIDINGWINDOW:5:25 MINLEN:80. The quality-filtered reads were aligned to the human reference genome (GRCh38) using Hisat2 v2.2.1[120]. A count table was then generated using featureCounts v2.0.1[121].

As sequencing was conducted in multiple batches, batch correction was applied to the count table using Combat_seq v3.50.0[122]. PCA was used to assess if the batch effect had been mitigated following batch correction. Once batch-corrected, genes were filtered to consider only those present in at least 50% of samples. PCA was performed using the *prcomp* function applied to the filtered and variance-stabilized transformed counts, produced by *DESeq2 v1.44.0*[123]. A Wilcoxon test was used to identify PCs significantly associated with relapse status. Differential expression analysis was also conducted using the R package *DESeq2*. Any patient features associated with relapse-associated PCs were included in the DESeq design. To identify DEGs, we used a Wald test and considered the remission group as the reference level. All reported *p*-values were adjusted for multiple testing using the Benjamini–Hochberg correction, and only those genes with an adjusted *p*-value < 0.05 were considered differentially expressed.

Analysis of associations between bacteria and host transcriptome

A lasso penalized regression approach, based on code provided by Priya et al.[124], was implemented to identify significant associations between bacterial abundance and human transcripts. For this analysis, transcript counts (after batch correction, see *Transcriptome analysis* section) were filtered and normalized using *DaMiR.normalization* function from DaMiRseq[125] package version 2.14, using a minimum threshold for transcript counts of 10, with 70% of samples meeting the threshold requirement. The minimum coefficient of variation considered was 5. The normalisation function used was *vst* (variance stabilising transformation). Bacterial abundances (calculated as detailed in the Microbiota *sequencing and analysis* section) were log10-transformed (adding a pseudocount of 1 to replace zeroes). The analysis was conducted separately for relapse and remission groups. *p*-values, determining the significance of each association, were then corrected on a single gene basis, taking into account only ASVs, differentially expressed between relapse and remission. Functional analysis was conducted separately for each set of transcripts associated with a single ASV, using the MSigDb C2 collection database (https://www.gsea-msigdb.org/gsea/msigdb/collection_details.jsp#C2) as a reference for human pathways. Fisher's exact test was a statistical method used, with a set of all tested genes as a background. To reduce the number of overlapping pathways, function *cluster_strings* from clustringr[126] R package version 1.0 was applied to lists of genes present in pathways, with a maximum distance set to 5. This analysis was complemented by functional analysis of human transcripts altered between relapse and remission. Given the large contribution of sampling sites to host transcriptome variation, this analysis was conducted on pooled results, as well as separately for each biopsy type. Since modest numbers of transcripts were found to be differentially expressed, Gene set enrichment analysis (GSEA) was used instead of Fisher's exact test. This method, which informs of pathways overrepresented towards the top or the bottom of the ranked gene list, was implemented in the reactomePA package version 1.4.2. The following metric was used to rank genes:

$$Score = \log_2 FoldChange \times (-\log_{10} pvalue)$$

This resulted in genes overexpressed in the future relapse group being near the top of the list and those overexpressed in the future remission group being placed closer to the bottom.

Bacteria growth

Both *Veillonella* species, *V. dispar* (DSM 20735) and *V. parvula* (DSM 2008), were purchased from the Leibniz Institute, DSMZ-German Collection of Microorganisms and Cell Cultures. *V. parvula* (DSM 2008) strain was originally isolated from the intestinal tract, while *V. dispar* (DSM 20735) originated in the human mouth. The bacterial strains were cultured anaerobically in brain heart infusion (BHI) medium supplemented with 1% Lactose and 0.05% L-Cysteine hydrochloride (Merck, Darmstadt, Germany). To collect the conditioned medium, each bacterium was grown to stationary phase, as previously described[127], followed by collection of the conditioned medium, centrifugation at maximum speed (3220 rpm), filter sterilization with 0.22 μm filters (Fisher Scientific, UK), and aliquoting and storing at −80 °C until further analysis.

Cell lines and co-culture

The human colorectal epithelial cell lines, HT29 (American Type Culture Collection, ATCC) and HCT116-Dual™ cells (reporter cell line for NF-κB and & IRF pathways, Invivogen, Toulouse, France) were cultured in McCoy's 5A and DMEM medium (Merck), respectively. Cells were seeded (2×10^5 cells/mL) in 96-well plates, cultured with full medium supplemented with 10% foetal bovine serum for 24 h, followed by culture in serum-free medium overnight, and the addition of the different stimuli as described below.

For co-culture experiments in HT29 and HCT116-Dual™ Cells, a 1:10 dilution of the *Veillonella* conditioned media was added for 24 h, with the cell supernatants and collected for further analysis.

For co-culture experiments and IBD drugs, HT29 and HCT116-Dual™ Cells were pre-treated for 1 h with the IBD drugs, Tofacitinib (10μM, Merck), Methylprednisolone (10 mM, Merck), or Sulfasalazine (500 μg/mL, TOCRIS), followed by a further 24 h co-culture with *Veillonella* conditioned media. The positive controls recombinant-IL-1β (Preprotech/ThermoFisher) and Poly(dA:dT)/LyoVec™ (Invivogen) were used for the reporter HCT116 Dual cells. In all experiments, untreated and BHI broth (dilution 1:10) were used as negative controls. At the end of the experiment, cell supernatants were collected for further analysis. For the HCT116 Dual Cells, the addition of Quantiblu reagent, absorbance at 630 nm, revealed NFκB activation, while the addition of Quantiluc reagent revealed IRF-activation, as per the manufacturer's instructions. Epithelial cell secretion of IL-8, CXCL10, CCL20, and IL-1β was quantified on collected supernatants by ELISA (DuoSets, Biotechne, Minneapolis, MN, USA), and the CyQUANT™ LDH Cytotoxicity Assay Kit (Invitrogen, Waltham, MA, USA) was used to measure cell viability, following the manufacturer's instructions, respectively.

Mouse strain colonization

IL10−/− mice (B6.129P2-Il10tm1Cgn/J) were acquired from Charles River Laboratories, UK, and bred at the Bioscience Service Unit (BSU)-Annex at University College Cork. Mice were housed in individually ventilated cages (IVCs, OptiMICE), in a controlled environment (20–22 °C, 12 h light:dark cycle) and given food (2018 Teklad Global 18% Protein Rodent Diets (2018S, Envigo, UK) and water *ad libitum*.

The experimental design for this study followed a recently published protocol from our group, in which $IL10^{-/-}$ mice received streptomycin (5 g/L) for 24 h prior to oral gavage with the IBD-associated pathobiont adherent-invasive *Escherichia coli* (AIEC)[35]. Microbiota analysis three days post-antibiotic treatment revealed a composition similar to that observed at the end of the study, i.e., 17 days later, indicating rapid microbial recovery and limited long-term depletion. Thus, in this study, the PBS control group—receiving no bacterial gavage—was not considered to be heavily microbiota-depleted at the time of analysis. $IL10^{-/-}$ mice (males, weight over 25 g, age 8–18 weeks) were given streptomycin (5 g/L) in drinking water *ad libitum* for 24 h at day −4, followed by oral gavage with *V. dispar* and *V. parvula* solution

(10^8–10^9 CFU/mL) starting from day -3 for three consecutive days[35]. From day 0 onwards, mice were gavaged with bacterial solutions every second day until day 13. Mice gavaged with PBS at the same time-points as the *Veillonella* spp, served as negative control. Due to technical issues, two mice had to be excluded from this group, resulting in a final sample size of $n = 3$. However, consistent with prior studies from our laboratory using an identical experimental design, colon weights measured in PBS-treated IL10$^{-/-}$ mice from two independent cohorts[35] closely matched the values observed in the current study. These consistent findings support the reliability of our data despite the reduced sample size in the PBS group. In a second experiment, pre-colonized *V. parvula* IL10$^{-/-}$ mice were fed 100ppm piroxicam for 5 days and euthanized at day 6[35]. Faeces were collected on days 1, 4, 9, and 13 post-start of *Veillonella* spp gavage, plated on *Veillonella* Agar Base medium (HiMedia M416[128]) supplemented with Vancomycin (7.5 µg/mL), and colonies counted at 72–96 h in an anaerobic chamber at 37 °C. Levels of colonization were comparable between groups inoculated with the two *Veillonella* strains; however, low-level colonization by *Veillonella spp*. was also detected in the PBS control group. This likely reflects the use of *Veillonella* Agar Base medium (HiMedia M416[128]), which supports the growth of *Veillonella* species broadly and does not discriminate between strains. At the end of the studies, tissue (colon, caecum, and spleen), blood, faeces, and caecal content were collected. Colon length and weights of the colon, caecum, and spleen were measured. Animals were monitored daily for body weight, stool consistency, fur texture, rectal prolapse, and posture as previously reported in ref. 35. Colon tissue samples were homogenized and examined for mKC (IL-8 homologue) expression by ELISA as reported[35].

ML analysis

ML analysis was conducted to identify combinations of data types that best predict relapse 6 months after entering remission using baseline data. In this analysis, the three omics types outlined earlier (gut microbiome, host transcriptomics, and host epigenetics) were considered along with host genotype data, a dataset for which we did not have sufficient power to analyse on its own (i.e., GWAS). We also examined the effect of including certain patient characteristics in our training data, specifically, patient age (in months), mode of feeding (breast or bottle fed), and mode of delivery (c-section or vaginal birth). Models were trained using individual omics types, as well as multi-omics combinations, which were combined using a simple concatenation approach. To identify those combinations with the best performance, we initially trained models using only those samples with full coverage across all omics types ($n = 100$), as this ensured consistency in the underlying training samples and avoided the difficulty of handling large quantities of missing data.

Genotype data, used in the analysis, were processed with the following quality control steps: (1) removal of variants with missing genotype calls. (2) Removal of variants in which MAF (minor allele frequency) is smaller than 0.02, (3) removal of variants which are in linkage disequilibrium (r2 > 0.2), and (4) removal of variants in violation of Hardy-Weinberg equilibrium (p-value < 10E-6). All the steps were executed using plink version 1.9[129].

To reduce the dimensionality and complexity of individual omics types, several feature selection steps were performed. Microbial taxa were considered if present in more than 10% of samples. Host RNASeq features with low abundance (<0.01%) and low prevalence (<10%) were removed. A centred log-ratio transformation was also applied to the RNASeq dataset prior to ML analysis to account for the compositional nature of the data. SNPs from the host genotype dataset were also removed if the most common allele was present in more than 80% of patients. Given the large number of CpG sites generated by the host methylation arrays, we merged individual sites into methylated regions using the *mCSEAdata* package v1.22.0[37]. We considered both promoter and gene body regions, and CpG sites were determined to be in a region using the annotation R packages: IlluminaHumanMethylation450kanno.ilmn12.hg19 and IlluminaHumanMethylationEPICanno.ilm10b2.hg19. See Martorell-Marugán et al.[37] for further details on the definition of promoter and gene body regions. For both datasets, we used the mean methylation value across CpG sites to represent the methylation of that region and trained models using both beta values and *M*-values to assess any potential differences in performance. Following this initial preprocessing, a total of 280 microbial features, 11,302 gene expression features, 34,946 genotype features, 22,247 methylated gene body features, and 24,994 methylation promoter regions were considered when samples were pooled from each biopsy site. Feature sets for individual biopsy sites following initial preprocessing after provided in Supplementary Data 12.

Given the limited sample size and lack of an external validation set, we conducted a nested cross-validation performance assessment routine. The outer loop consisted of a Leave-One-Out (LOO) CV, and the held-out sample was used to assess the performance only and for no other purpose. In the case where more than one sample existed for a particular subject, the additional samples were also excluded from the training data. During the internal model development stage, additional feature selection steps were applied to the training data to further reduce the dimensionality of the datasets and remove potentially redundant features. First, we removed any features with near-zero variance. Then, a Mann–Whitney *U*-test (numeric features) or a Chi-Square test (categorical features) was used to identify any potentially informative features with respect to our outcome of interest. When a model was being trained on pooled samples (all biopsy sites together), these tests were performed for each site independently. Those features found to be statistically significant ($p < 0.05$) were selected to be included in the training dataset.

Our ML analysis was conducted using the decision tree-based boosting algorithm XGBoost[130]. The inner loop of our nested CV was used to tune model hyperparameters, where we performed a randomized search based on a 5-fold CV (or Grouped 5-fold CV for pooled samples). In our analysis, we tuned the following hyperparameters: *max_depth, colsample_bytree, subsample*, and *reg_alpha*. Once optimal hyperparameters were selected, we created an ensemble of 10 XGBoost models by splitting the training data into 10 folds. Nine folds were used to train each model, and the remaining fold was used as a model checkpoint and for early stopping (*early_stopping_rounds* = 20). That is, if performance on the held-out fold did not improve after 20 rounds, training was stopped, and the model was saved. Such early stopping helps prevent the model from overfitting to the training data[131]. This procedure was repeated for each combination of folds. Once the 10 models were trained, the average prediction for the held-out test sample was stored and used to assess the performance of our approach.

For all models, we assessed performance using the area under the ROC curve (AUC) metric. Performance was evaluated for the held-out *test set*, and the *validation set* performance was also estimated using predictions from the inner loop of our nested CV process. By comparing the test and validation performance, it was possible to assess the stability and generalizability of our approach. Furthermore, we extracted feature importance values from models to assess which features were playing the biggest role in the model predictions. These importance values were averaged across each iteration of the CV and were normalized by the number of times it was found to have non-zero importance in the ensemble. All ML analysis was conducted using the following Python packages: *xgboost v1.7.6, pandas v2.0.3, scikit-learn v1.3.0, numpy v1.25.0*, and *scipy v1.11.1*.

Once the top-performing models were identified, error analysis was conducted to assess whether the errors made by the model related to specific metadata features. To do this, we first calculated the optimal classification threshold using the geometric-mean (G-mean) approach.

The G-mean is defined as the square root of the product of sensitivity (TPR—true positive rate) and specificity (TNR—true negative rate).

$$G-mean=\sqrt{Sensitivity\times Specificity}$$

The threshold for classification is selected such that the G-mean value is maximized. As a result, this threshold represents the best trade-off between sensitivity and specificity, ensuring balanced performance for both classes. Once the optimal threshold was selected, a Fisher's exact test was used to compare the proportion of misclassifications between various metadata features. Finally, as our initial analysis was conducted using only those samples with complete data across all four omics, additional samples were available for many of the top-performing data combinations. Additional samples, from patients not previously included in the ML analysis, were incorporated into the ML pipeline to further evaluate the performance. This involved repeating our nested CV performance assessment routine (described previously), but this time using samples with complete data plus any additional samples available for the data combination under consideration. We repeated the same intermediate steps in terms of feature selection, hyperparameter tuning, and model training, as before, and evaluated the performance using the AUC metric. This was repeated for the top 3 pooled sample models and the top 3 models from each biopsy site.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

16S rRNA sequencing data are available in the Sequence Read Archive (SRA, https://www.ncbi.nlm.nih.gov/sra) under accession number PRJNA1181771. Due to privacy restrictions, host-related data (RNASeq and epigenetic data) are available from the corresponding author for non-commercial purposes.

Code availability

Code used to generate results in this publication is available at https://doi.org/10.5281/zenodo.15433759.

References

1. Levine, A. et al. Pediatric modification of the Montreal classification for inflammatory bowel disease: the Paris classification. *Inflamm. Bowel Dis.* **17**, 1314–1321 (2011).
2. Fell, J. M. et al. Management of ulcerative colitis. *Arch. Dis. Child.* **101**, 469–474 (2016).
3. Aloi, M. et al. Presenting features and disease course of pediatric ulcerative colitis. *J. Crohns Colitis* **7**, e509–e515 (2013).
4. Dhaliwal, J. et al. Phenotypic variation in paediatric inflammatory bowel disease by age: a multicentre prospective inception cohort study of the Canadian children IBD network. *J. Crohns Colitis* **14**, 445–454 (2020).
5. Olén, O. et al. Increased mortality of patients with childhood-onset inflammatory bowel diseases, compared with the general population. *Gastroenterology* **156**, 614–622 (2019).
6. Bitton, A., Vutcovici, M., Sewitch, M., Suissa, S. & Brassard, P. Mortality trends in Crohn's disease and ulcerative colitis: a population-based study in Québec, Canada. *Inflamm. Bowel Dis.* **22**, 416–423 (2016).
7. Kuenzig, M. E. et al. Twenty-first century trends in the global epidemiology of pediatric-onset inflammatory bowel disease: systematic review. *Gastroenterology* **162**, 1147–1159.e4 (2022).
8. Halfvarson, J. et al. Age determines the risk of familial inflammatory bowel disease—a nationwide study. *Aliment. Pharmacol. Ther.* **56**, 491–500 (2022).
9. Ostrowski, J. et al. Genetic architecture differences between pediatric and adult-onset inflammatory bowel diseases in the Polish population. *Sci. Rep.* **6**, 39831 (2016).
10. Knights, D., Lassen, K. G. & Xavier, R. J. Advances in inflammatory bowel disease pathogenesis: linking host genetics and the microbiome. *Gut* **62**, 1505–1510 (2013).
11. Hansen, R. et al. Microbiota of de-novo pediatric IBD: increased Faecalibacterium prausnitzii and reduced bacterial diversity in Crohn's but not in ulcerative colitis. *Am. J. Gastroenterol.* **107**, 1913–1922 (2012).
12. Alipour, M. et al. Mucosal barrier depletion and loss of bacterial diversity are primary abnormalities in paediatric ulcerative colitis. *J. Crohns Colitis* **10**, 462–471 (2016).
13. Clooney, A. G. et al. Ranking microbiome variance in inflammatory bowel disease: a large longitudinal intercontinental study. *Gut* **70**, 499–510 (2021).
14. Ryan, F. J. et al. Colonic microbiota is associated with inflammation and host epigenomic alterations in inflammatory bowel disease. *Nat. Commun.* **11**, 1512 (2020).
15. Schirmer, M. et al. Compositional and temporal changes in the gut microbiome of pediatric ulcerative colitis patients are linked to disease course. *Cell Host Microbe* **24**, 600–610.e4 (2018).
16. Fitzgerald, R. S., Sanderson, I. R. & Claesson, M. J. Paediatric inflammatory bowel disease and its relationship with the microbiome. *Microb. Ecol.* **82**, 833–844 (2021).
17. Russell, R. K. et al. Contemporary outcomes for ulcerative colitis inpatients admitted to pediatric hospitals in the United Kingdom. *Inflamm. Bowel Dis.* **19**, 1434–1440 (2013).
18. Turner, D. et al. Management of pediatric ulcerative colitis: joint ECCO and ESPGHAN evidence-based consensus guidelines. *J. Pediatr. Gastroenterol. Nutr.* **55**, 340–361 (2012).
19. Turner, D. et al. Management of paediatric ulcerative colitis, part 2: acute severe colitis-an evidence-based consensus guideline from the European Crohn's and Colitis Organization and the European Society of Paediatric Gastroenterology, Hepatology and Nutrition. *J. Pediatr. Gastroenterol. Nutr.* **67**, 292–310 (2018).
20. Turner, D. et al. Faecal calprotectin, lactoferrin, M2-pyruvate kinase and S100A12 in severe ulcerative colitis: a prospective multicentre comparison of predicting outcomes and monitoring response. *Gut* **59**, 1207–1212 (2010).
21. Schechter, A. et al. Early endoscopic, laboratory and clinical predictors of poor disease course in paediatric ulcerative colitis. *Gut* **64**, 580–588 (2015).
22. Lee, J. C. et al. Human SNP links differential outcomes in inflammatory and infectious disease to a FOXO3-regulated pathway. *Cell* **155**, 57–69 (2013).
23. Gasparetto, M. et al. Transcription and DNA methylation patterns of blood derived CD8 + T cells are associated with age and inflammatory bowel disease but do not predict prognosis. *Gastroenterology* **160**, 232–244.e7 (2021).
24. Lee, J. C. et al. Gene expression profiling of CD8 + T cells predicts prognosis in patients with Crohn disease and ulcerative colitis. *J. Clin. Invest.* **121**, 4170–4179 (2011).
25. Hyams, J. S. et al. Clinical and biological predictors of response to standardised paediatric colitis therapy (PROTECT): a multicentre inception cohort study. *Lancet* **393**, 1708–1720 (2019).
26. Krishnakumar, C. et al. Early Change in fecal calprotectin predicts one-year outcome in children newly diagnosed with ulcerative colitis. *J. Pediatr. Gastroenterol. Nutr.* **74**, 72–78 (2022).
27. Cleynen, I. et al. Inherited determinants of Crohn's disease and ulcerative colitis phenotypes: a genetic association study. *Lancet Lond. Engl.* **387**, 156–167 (2016).
28. Jian, C., Luukkonen, P., Yki-Järvinen, H., Salonen, A. & Korpela, K. Quantitative PCR provides a simple and accessible method for quantitative microbiota profiling. *PLoS One* **15**, e0227285 (2020).

29. Rivière, A., Selak, M., Lantin, D., Leroy, F. & De Vuyst, L. Bifidobacteria and butyrate-producing colon bacteria: importance and strategies for their stimulation in the human gut. *Front. Microbiol.* **7**, 979 (2016).
30. Tailford, L. E., Crost, E. H., Kavanaugh, D. & Juge, N. Mucin glycan foraging in the human gut microbiome. *Front. Genet.* **6**, 81 (2015).
31. Aniwan, S., Tremaine, W. J., Raffals, L. E., Kane, S. V. & Loftus, E. V. Antibiotic use and new-onset inflammatory bowel disease in Olmsted County, Minnesota: a population-based case-control study. *J. Crohns Colitis* **12**, 137–144 (2018).
32. Faye, A. S. et al. Antibiotic use as a risk factor for inflammatory bowel disease across the ages: a population-based cohort study. *Gut* **72**, 663–670 (2023).
33. Duan, R., Zhang, C., Li, G., Li, J. & Duan, L. Antibiotic exposure and risk of new-onset inflammatory bowel disease: a systematic review and dose-response meta-analysis. *Clin. Gastroenterol. Hepatol. Off. Clin. Pract. J. Am. Gastroenterol. Assoc.* **23**, 45–58.e15 (2025).
34. Knights, D. et al. Rethinking "Enterotypes". *Cell Host Microbe* **16**, 433–437 (2014).
35. Singh, R. et al. An IBD-associated pathobiont synergises with NSAID to promote colitis which is blocked by NLRP3 inflammasome and Caspase-8 inhibitors. *Gut Microbes* **15**, 2163838 (2023).
36. Leenen, F. A. D., Muller, C. P. & Turner, J. D. DNA methylation: Conducting the orchestra from exposure to phenotype?. *Clin. Epigenetics* **8**, 92 (2016).
37. Martorell-Marugán, J., González-Rumayor, V. & Carmona-Sáez, P. mCSEA: detecting subtle differentially methylated regions. *Bioinformatics* **35**, 3257–3262 (2019).
38. Shi, F. et al. Dysregulated Tim-3 expression and its correlation with imbalanced CD4 helper T cell function in ulcerative colitis. *Clin. Immunol.* **145**, 230–240 (2012).
39. Vainer, B. Intercellular adhesion molecule-1 (ICAM-1) in ulcerative colitis: presence, visualization, and significance. *Inflamm. Res. Off. J. Eur. Histamine Res. Soc. Al* **54**, 313–327 (2005).
40. UniProtKB. GAPT—Protein GAPT—*Homo sapiens* (Human). UniProt. https://www.uniprot.org/uniprotkb/Q8N292/entry#function Accessed 26 August 2024.
41. Bertocci, B. et al. Klhl6 deficiency impairs transitional B cell survival and differentiation. *J. Immunol.* **199**, 2408–2420 (2017).
42. Borcard, D., Gillet, F. & Legendre, P. *Numerical Ecology with R* (Springer International Publishing, Cham, 2018); https://doi.org/10.1007/978-3-319-71404-2.
43. Pilla-Moffett, D., Barber, M. F., Taylor, G. A. & Coers, J. Interferon-inducible GTPases in host resistance, inflammation and disease. *J. Mol. Biol.* **428**, 3495–3513 (2016).
44. Lin, X.-W. et al. WW domain containing E3 ubiquitin protein ligase 1 (WWP1) negatively regulates TLR4-mediated TNF-α and IL-6 production by proteasomal degradation of TNF receptor associated factor 6 (TRAF6). *PLoS One* **8**, e67633 (2013).
45. Schoenborn, J. R. & Wilson, C. B. Regulation of interferon-γ during innate and adaptive immune responses. In *Advances in Immunology*, Vol. 96, 41–101 (Academic Press, 2007).
46. Tu, Q. et al. Prokineticin 2 promotes macrophages-mediated antibacterial host defense against bacterial pneumonia. *Int. J. Infect. Dis.* **125**, 103–113 (2022).
47. Subramanian, A. et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. USA* **102**, 15545–15550 (2005).
48. Hu, S. et al. Mucosal host–microbe interactions associate with clinical phenotypes in inflammatory bowel disease. *Nat. Commun.* **15**, 1470 (2024).
49. Ma, B. et al. Diagnostic classification of cancers using extreme gradient boosting algorithm and multi-omics data. *Comput. Biol. Med.* **121**, 103761 (2020).
50. Hou, X. et al. The transcriptional risk scores for kidney renal clear cell carcinoma using XGBoost and multiple omics data. *Math. Biosci. Eng.* **20**, 11676–11687 (2023).
51. Turner, D. et al. Development, validation, and evaluation of a pediatric ulcerative colitis activity index: a prospective multicenter study. *Gastroenterology* **133**, 423–432 (2007).
52. Borren, N. Z. & Ananthakrishnan, A. N. Precision medicine: How multi-'omics will take shape the future of IBD? *Curr. Opin. Gastroenterol.* **38**, 382–387 (2022).
53. Qiu, X., Zhang, M., Yang, X., Hong, N. & Yu, C. *Faecalibacterium prausnitzii* upregulates regulatory T cells and anti-inflammatory cytokines in treating TNBS-induced colitis. *J. Crohns Colitis* **7**, e558–e568 (2013).
54. Machiels, K. et al. A decrease of the butyrate-producing species *Roseburia hominis* and *Faecalibacterium prausnitzii* defines dysbiosis in patients with ulcerative colitis. *Gut* **63**, 1275–1283 (2014).
55. Franzosa, E. A. et al. Gut microbiome structure and metabolic activity in inflammatory bowel disease. *Nat. Microbiol.* **4**, 293–305 (2019).
56. Png, C. W. et al. Mucolytic bacteria with increased prevalence in IBD mucosa augmentIn vitroutilization of mucin by other bacteria. *Off. J. Am. Coll. Gastroenterol. ACG* **105**, 2420 (2010).
57. Kitamoto, S., Nagao-Kitamoto, H., Hein, R., Schmidt, T. M. & Kamada, N. The bacterial connection between the oral cavity and the gut diseases. *J. Dent. Res.* **99**, 1021–1029 (2020).
58. Michail, S. et al. Alterations in the gut microbiome of children with severe ulcerative colitis. *Inflamm. Bowel Dis.* **18**, 1799–1808 (2012).
59. Geva-Zatorsky, N. et al. Mining the human gut microbiota for immunomodulatory organisms. *Cell* **168**, 928–943.e11 (2017).
60. Zhao, N. et al. Role of Interleukin-22 in ulcerative colitis. *Biomed. Pharmacother.* **159**, 114273 (2023).
61. Rojas-Tapias, D. F. et al. Inflammation-associated nitrate facilitates ectopic colonization of oral bacterium *Veillonella parvula* in the intestine. *Nat. Microbiol.* **7**, 1673–1685 (2022).
62. Schirmer, M. et al. Linking microbial genes to plasma and stool metabolites uncovers host-microbial interactions underlying ulcerative colitis disease course. *Cell Host Microbe* **32**, 209–226.e7 (2024).
63. van den Bogert, B., Meijerink, M., Zoetendal, E. G., Wells, J. M. & Kleerebezem, M. Immunomodulatory properties of *Streptococcus* and *Veillonella* isolates from the human small intestine microbiota. *PLoS One* **9**, e114277 (2014).
64. Zhan, Z. et al. Overabundance of *Veillonella parvula* promotes intestinal inflammation by activating macrophages via LPS-TLR4 pathway. *Cell Death Discov* **8**, 251 (2022).
65. Ke, L. et al. Investigation on the pathological mechanism of frequent exacerbators with chronic obstructive pulmonary disease based on the characteristics of respiratory flora. *Front. Med.* **8**, 816802 (2022).
66. Giacomini, J. J., Torres-Morales, J., Dewhirst, F. E., Borisy, G. G. & Mark Welch, J. L. Site specialization of human oral *Veillonella* Species. *Microbiol. Spectr.* **11**, e04042–22 (2023).
67. Haberman, Y. et al. Ulcerative colitis mucosal transcriptomes reveal mitochondriopathy and personalized mechanisms underlying disease severity and treatment response. *Nat. Commun.* **10**, 38 (2019).
68. Howell, K. J. 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* **154**, 585–598 (2018).
69. Ito, R. et al. Interferon-gamma is causatively involved in experimental inflammatory bowel disease in mice. *Clin. Exp. Immunol.* **146**, 330–338 (2006).
70. Yang, S. et al. Interferon-γ-induced intestinal epithelial barrier dysfunction by NF-κB/HIF-1α pathway. *J. Interferon Cytokine Res. Off. J. Int. Soc. Interferon Cytokine Res.* **34**, 195–203 (2014).

71. Wang, F. et al. Interferon-γ and tumor necrosis factor-α synergize to induce intestinal epithelial barrier dysfunction by up-regulating myosin light chain kinase expression. *Am. J. Pathol.* **166**, 409–419 (2005).
72. Naschberger, E. et al. Analysis of the interferon-γ-induced secretome of intestinal endothelial cells: putative impact on epithelial barrier dysfunction in IBD. *Front. Cell Dev. Biol.* **11**, 1213383 (2023).
73. Langer, V. et al. IFN-**γ** drives inflammatory bowel disease pathogenesis through VE-cadherin–directed vascular barrier disruption. *J. Clin. Invest.* **129**, 4691–4707 (2019).
74. Brasseit, J. et al. Divergent Roles of Interferon-γ and Innate Lymphoid Cells in Innate and Adaptive Immune Cell-Mediated Intestinal Inflammation. *Front. Immunol.* **9**, 23 (2018).
75. Li, J. et al. Profiles of lamina propria T helper cell subsets discriminate between ulcerative colitis and Crohn's disease. *Inflamm. Bowel Dis.* **22**, 1779–1792 (2016).
76. Hu, X., Li, W.-P., Meng, C. & Ivashkiv, L. B. Inhibition of IFN-γ signaling by glucocorticoids1. *J. Immunol.* **170**, 4833–4839 (2003).
77. Musch, E. et al. Interferon-β–1a for the treatment of steroid-refractory ulcerative colitis: a randomized, double-blind, placebo-controlled trial. *Clin. Gastroenterol. Hepatol.* **3**, 581–586 (2005).
78. Mannon, P. J. et al. Suppression of inflammation in ulcerative colitis by interferon-β–1a is accompanied by inhibition of IL-13 production. *Gut* **60**, 449–455 (2011).
79. Katakura, K. et al. Toll-like receptor 9–induced type I IFN protects mice from experimental colitis. *J. Clin. Invest.* **115**, 695–702 (2005).
80. McElrath, C. et al. Critical role of interferons in gastrointestinal injury repair. *Nat. Commun.* **12**, 2624 (2021).
81. Carow, B. & Rottenberg, M. E. SOCS3, a major regulator of infection and inflammation. *Front. Immunol.* **5**, 58 (2014).
82. Bereshchenko, O., Migliorati, G., Bruscoli, S. & Riccardi, C. Glucocorticoid-induced leucine zipper: a novel anti-inflammatory molecule. *Front. Pharmacol.* **10**, 308 (2019).
83. Li, S., Ning, L.-G., Lou, X.-H. & Xu, G.-Q. Necroptosis in inflammatory bowel disease and other intestinal diseases. *World J. Clin. Cases* **6**, 745–752 (2018).
84. Pappa, A. et al. Pediatric IBD patients show medication and disease activity dependent changes in NK cell and CD4 memory T cell populations. *Front. Pediatr.* **11**, 1123873 (2023).
85. Chojnacki, A. K. et al. Tissue imaging reveals disruption of epithelial mitochondrial networks and loss of mitochondria-associated cytochrome-C in inflamed human and murine colon. *Mitochondrion* **68**, 44–59 (2023).
86. Clarkston, K. et al. Targeted assessment of mucosal immune gene expression predicts clinical outcomes in children with ulcerative colitis. *J. Crohns Colitis* **16**, 1735–1750 (2022).
87. Duranti, S. et al. Elucidating the gut microbiome of ulcerative colitis: Bifidobacteria as novel microbial biomarkers. *FEMS Microbiol. Ecol.* **92**, fiw191 (2016).
88. Sang, L.-X. et al. Remission induction and maintenance effect of probiotics on ulcerative colitis: A meta-analysis. *World J. Gastroenterol. WJG* **16**, 1908–1915 (2010).
89. Blesl, A. et al. Prediction of response to systemic corticosteroids in active UC by microbial composition—a prospective multicenter study. *Inflamm. Bowel Dis.* **30**, 9–19 (2024).
90. Lee, J. W. J. et al. Multi-omics reveal microbial determinants impacting responses to biologic therapies in inflammatory bowel disease. *Cell Host Microbe* **29**, 1294–1304.e4 (2021).
91. Ventin-Holmberg, R. et al. Bacterial and fungal profiles as markers of infliximab drug response in inflammatory bowel disease. *J. Crohns Colitis* **15**, 1019–1031 (2021).
92. Jones, P. A. Functions of DNA methylation: islands, start sites, gene bodies and beyond. *Nat. Rev. Genet.* **13**, 484–492 (2012).
93. Wang, Q. et al. Gene body methylation in cancer: molecular mechanisms and clinical applications. *Clin. Epigenetics* **14**, 154 (2022).
94. Yang, X. et al. Gene body methylation can alter gene expression and is a therapeutic target in cancer. *Cancer Cell* **26**, 577–590 (2014).
95. Turner, D. et al. Management of paediatric ulcerative colitis, part 1: ambulatory care-an evidence-based guideline from European Crohn's and Colitis Organization and European Society of Paediatric Gastroenterology, Hepatology and Nutrition. *J. Pediatr. Gastroenterol. Nutr.* **67**, 257–291 (2018).
96. Klindworth, A. et al. Evaluation of general 16S ribosomal RNA gene PCR primers for classical and next-generation sequencing-based diversity studies. *Nucleic Acids Res.* **41**, e1 (2013).
97. Patel, S. et al. Differential analysis of longitudinal methicillin-resistant *Staphylococcus aureus* colonization in relation to microbial shifts in the nasal microbiome of neonatal piglets. *mSystems* **6**, e0015221 (2021).
98. Babraham Bioinformatics. *FastQC—A Quality Control Tool for High Throughput Sequence Data*. https://www.bioinformatics.babraham.ac.uk/projects/fastqc/ (2018).
99. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120 (2014).
100. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal* **17**, 10–12 (2011).
101. Callahan, B. J. et al. DADA2: High-resolution sample inference from Illumina amplicon data. *Nat. Methods* **13**, 581–583 (2016).
102. Davis, N. M., Proctor, D. M., Holmes, S. P., Relman, D. A. & Callahan, B. J. Simple statistical identification and removal of contaminant sequences in marker-gene and metagenomics data. *Microbiome* **6**, 226 (2018).
103. Schloss, P. D. et al. Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Appl. Environ. Microbiol.* **75**, 7537–7541 (2009).
104. Quast, C. et al. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Res.* **41**, D590–D596 (2013).
105. Allard, G., Ryan, F. J., Jeffery, I. B. & Claesson, M. J. SPINGO: a rapid species-classifier for microbial amplicon sequences. *BMC Bioinformatics* **16**, 324 (2015).
106. Douglas, G. M. et al. PICRUSt2 for prediction of metagenome functions. *Nat. Biotechnol.* **38**, 685–688 (2020).
107. Dixon, P. & VEGAN, A. Package of R functions for community ecology. *J. Veg. Sci.* **14**, 927–930 (2003).
108. Zhang, X. & Yi, N. NBZIMM: negative binomial and zero-inflated mixed models, with application to microbiome/metagenomics data analysis. *BMC Bioinformatics* **21**, 488 (2020).
109. Bork Group. VpThemAll. *GitLab* https://git.embl.de/grp-bork/vpthemall (2021).
110. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. Ser. B Methodol.* **57**, 289–300 (1995).
111. Langfelder, P., Zhang, B. & Horvath, S. Defining clusters from a hierarchical cluster tree: the dynamic tree cut package for R. *Bioinformatics* **24**, 719–720 (2008).
112. Cáceres, M. D. & Legendre, P. Associations between species and groups of sites: indices and statistical inference. *Ecology* **90**, 3566–3574 (2009).
113. Aryee, M. J. et al. Minfi: a flexible and comprehensive Bioconductor package for the analysis of Infinium DNA methylation microarrays. *Bioinformatics* **30**, 1363–1369 (2014).

114. Benton, M. C. Illumina450K_filtering: a collection of resources to filter Illumina 450k and EPIC methylation arrays. Github https://github.com/sirselim/illumina450k_filtering (2016).
115. Pidsley, R. et al. Critical evaluation of the Illumina MethylationEPIC BeadChip microarray for whole-genome DNA methylation profiling. *Genome Biol* **17**, 208 (2016).
116. Chen, Y. et al. Discovery of cross-reactive probes and polymorphic CpGs in the Illumina Infinium HumanMethylation450 microarray. *Epigenetics* **8**, 203–209 (2013).
117. Teschendorff, A. E. et al. A beta-mixture quantile normalization method for correcting probe design bias in Illumina Infinium 450 k DNA methylation data. *Bioinforma. Oxf. Engl.* **29**, 189–196 (2013).
118. Johnson, W. E., Li, C. & Rabinovic, A. Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostat. Oxf. Engl.* **8**, 118–127 (2007).
119. Kuznetsova, A., Brockhoff, P. B. & Christensen, R. H. B. lmerTest package: tests in linear mixed effects models. *J. Stat. Softw.* **82**, 1–26 (2017).
120. Kim, D., Paggi, J. M., Park, C., Bennett, C. & Salzberg, S. L. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nat. Biotechnol.* **37**, 907–915 (2019).
121. Liao, Y., Smyth, G. K. & Shi, W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* **30**, 923–930 (2014).
122. Zhang, Y., Parmigiani, G. & Johnson, W. E. ComBat-seq: batch effect adjustment for RNA-seq count data. *NAR Genomics Bioinforma.* **2**, lqaa078 (2020).
123. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol* **15**, 550 (2014).
124. Priya, S. et al. Identification of shared and disease-specific host gene–microbiome associations across human diseases using multi-omic integration. *Nat. Microbiol.* **7**, 780–795 (2022).
125. Chiesa, M., Colombo, G. I. & Piacentini, L. DaMiRseq-an R/Bioconductor package for data mining of RNA-Seq data: normalization, feature selection and classification. *Bioinforma. Oxf. Engl.* **34**, 1416–1418 (2018).
126. Reznik, D. S. *dan-reznik/clustringr* (2024).
127. Saiz-Gonzalo, G. et al. Regulation of CEACAM family members by IBD-associated triggers in intestinal epithelial cells, their correlation to inflammation and relevance to IBD pathogenesis. *Front. Immunol.* **12**, 655960 (2021).
128. Rogosa, M., Fitzgerald, R. J., MacKintosh, M. E. & Beaman, A. J. Improved medium for selective isolation of veillonella. *J. Bacteriol.* **76**, 455–456 (1958).
129. Chang, C. C. et al. Second-generation PLINK: rising to the challenge of larger and richer datasets. *GigaScience* **4**, s13742–015-0047–8 (2015).
130. Chen, T. & Guestrin, C. XGBoost: a scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 785–794 (Association for Computing Machinery, New York, 2016); https://doi.org/10.1145/2939672.2939785.
131. Prechelt, L. Early stopping—But when? In *Neural Networks: Tricks of the Trade* 53–67 (Springer, Berlin, Heidelberg, 2012); https://doi.org/10.1007/978-3-642-35289-8_5.

Acknowledgements

This publication has emanated in part from research conducted with the financial support of Taighde Éireann – Research Ireland under grant number SFI/12/RC/2273_P2 (centre grant), 17/CDA/4765 (to M.J.C), and 18/CRT/6214 (to J.O.S), as well as the Crohn's and Colitis Foundation Litwin IBD Pioneers program (grant number 510264 to I.R.S). Maria Kulecka was partially funded by the INSPIRE COFUND Marie Skłodowska-Curie grant agreement no. 101034270. Patricia Ruiz-Limón was supported by a "Miguel Servet" postdoctoral contract (CP22/00096) by the Instituto de Salud Carlos III, Madrid, Spain, and co-funded by the European Union. We acknowledge and thank the support of the staff at the Biological Service Unit-Annex at University College Cork for the animal studies. We are grateful to Barts Health NHS Trust Children's Clinical Research facility, including the nursing staff and the clinical research fellows, Natalie Babb, Ramiya Kirupananthan, Natasha Thorn, and Kirn Sandhu, and to the children and their parents who consented to participate in the study.

Author contributions

M.J.C., I.R.S., S.M., and P.D. designed and managed the project. M.K., J.O.S., and R.F. performed the statistical analyses; I.R.S., P.D., N.M.C., and S.N. provided clinical expertise and facilitated the collection of samples. A.V., C.H., P.R.L., J.E., A.Z., and E.J.L.M. performed the laboratory analysis on clinical samples. M.V.M. and B.A.T. conducted experiments and analysis on animals. M.V.M. and R.S. conducted experiments and analysis on cell lines. M.K., J.O.S., S.M., I.R.S., and M.J.C. wrote the manuscript. J.E., A.T., and P.D. reviewed and edited the manuscript. M.J.C. and I.R.S. secured funding.

Competing interests

M.J.C. is a co-founder of SeqBiome Ltd. The remaining authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41467-025-62533-z.

Correspondence and requests for materials should be addressed to Marcus J. Claesson.

Peer review information *Nature Communications* thanks Kiana West, and the other, anonymous, reviewers for their contribution to the peer review of this work. A peer review file is available.