Tissue-dependent transcriptional and bacterial associations in primary sclerosing cholangitis-associated inflammatory bowel disease [version 1; peer review: awaiting peer review]

Nicholas E. Ilott1, Mastura Neyazi2,
Oxford Translational Gastroenterology Unit Investigators,
Carolina V. Arancibia-Cárcamo2, Fiona Powrie1,2, Alessandra Geremia2

1The Kennedy Institute of Rheumatology, Nuffield Department of Orthopaedics, Rheumatology and Musculoskeletal Sciences, University of Oxford, Oxford, OX3 7FY, UK
2Translational Gastroenterology Unit, Nuffield Department of Clinical Medicine, Experimental Medicine Division, and NIHR Oxford Biomedical Research Centre, Oxford University Hospitals NHS Foundation Trust, John Radcliffe Hospital, University of Oxford, Oxford, OX3 9DU, UK

Abstract
Background: Primary sclerosing cholangitis (PSC) is a disease of the bile duct and liver. However, patients frequently have co-morbidities including inflammatory bowel disease (IBD) and colorectal cancer. Colorectal cancer risk in patients with PSC-associated ulcerative colitis (PSC/UC) is elevated relative to patients with ulcerative colitis (UC) alone, reasons for which remain obscure. Further, clinical and immunological features, and involved intestinal sites differ between PSC/UC and UC. Understanding the molecular and microbial basis for differences in cancer risk between these two patient groups and how these differ across intestinal sites is important for the development of therapies to prevent colorectal cancer development in at-risk individuals.

Methods: We employed ribonucleic acid sequencing (RNA-seq) analysis of biopsy samples across three intestinal tissue locations (ileum, caecum and rectum) in patients with PSC/UC (n = 8), UC (n = 10) and healthy controls (n = 12) to determine tissue-dependent transcriptional alterations in PSC/UC. We also performed 16S ribosomal RNA (rRNA) amplicon sequencing to determine bacterial associations with PSC/UC and host-microbiome associations.

Results: Tissue-defining transcriptional signatures revealed that the ileum was enriched for genes involved in lipid and drug metabolism, the caecum for activated immune cells and the rectum for enteric neurogenesis. Transcriptional alterations relative to healthy control samples were largely shared between patients with PSC/UC or UC although were distinct across tissue locations. Nevertheless, we observed reduced expression of gamma-glutamyl transferase 1 (GGT1)
specifically in the ileum and caecum of patients with PSC/UC. Analysis of the bacterial component of the microbiome revealed high inter-individual variability of microbiome composition and little evidence for tissue-dependency. We observed a reduction in Parabacteroides relative abundance in the rectum of patients with PSC/UC.

**Conclusions:** The role of gamma-glutamyl transferase in maintaining the redox environment through the glutathione salvage pathway makes our observed alterations a potential pathway to PSC-associated colorectal cancer.

**Keywords**
Primary sclerosing cholangitis, IBD, transcriptomics, microbiome
**Introduction**

Primary sclerosing cholangitis (PSC) is a chronic progressive disorder of the hepatobiliary system characterised by inflammation, fibrosis, and stricturing of the intrahepatic and extrahepatic bile ducts. A common co-occurring disease in patients with PSC is inflammatory bowel disease (IBD), with a prevalence of 60–80%. While most frequently diagnosed as ulcerative colitis (UC), clinical features of colitis patients with PSC are somewhat distinct. These features include more frequent pancolitis or right-sided patchy inflammation, rectal sparing, “backwash ileitis” and quiescent course. Further, an important differential disease course between patients with PSC/UC and patients with UC is the fivefold increased risk of colorectal cancer in PSC/UC relative to UC. The reasons for this increased risk are currently unclear but in the context of colitis-associated cancer may involve pro-inflammatory immune phenotypes that promote cell growth and proliferation, reduced ability to clear deoxyribonucleic acid (DNA)-damaging reactive oxygen species, defective DNA-damage repair mechanisms or alterations to the gut microbiota that favour genotoxic species.

Clinical and disease course distinctions are important when considering a molecular and cellular basis for disease. Indeed, as a complex disease, genome-wide association studies (GWAS) have identified multiple genetic variants of small effect associated with PSC. The most highly associated risk variants are those that reside in the human leukocyte antigen (HLA) locus with non-HLA variants predominantly associated with adaptive immunity. Nevertheless, the genetic correlation between PSC/UC and UC is relatively low, suggesting distinct genetic architectures between these diseases. Furthermore, while immune phenotypes that differentiate PSC/UC from UC have not been extensively studied, and although broadly similar in major cell types, there is some evidence to suggest an increased frequency of innate lymphoid cells and interferon gamma (IFN-γ) producing CD4 T cells in the colon of patients with PSC/UC relative to UC. RNA-sequencing analysis of colonic biopsies has provided links between PSC/UC and increased expression of genes involved in bile acid homeostasis and decreased expression of genes involved in DNA damage repair relative to patients with UC. These data support alterations in liver-gut cross-talk that distinguishes colonic transcriptomes of patients with PSC/UC from those without.

The aforementioned cellular and molecular studies in PSC/UC have focussed on colonic samples. However, observed variation in tissue involvement between PSC/UC and UC provides an impetus to extend studies to both the large and the small intestine that may have distinct roles to play in pathogenesis. Indeed, the structure and function of the intestine varies across its length. The small intestine is home to a high density of absorptive enterocytes that promote nutrient uptake where the main role of the colon is to remove water from stool and accommodate a high density of microorganisms – the microbiome - that aid in digestion, peristalsis and immune function. Subtle and significant differences in cellular composition between small intestinal and colonic sites in terms of epithelial and immune cell types supports a requirement to study isolated and integrated effects on disease across tissue sites.

There is growing evidence for a role of the gut microbiome in PSC pathogenesis. Indeed, multiple microbiome association studies in both stool and the intestinal mucosa have identified alterations in the bacterial component of the microbiome in PSC and/or PSC/UC. Increased relative abundance of Enterococcus, Veillonella and Streptococcus genera have been reported in multiple studies, with Klebsiella pneumoniae-harbouring microbiomes from patients with PSC/UC being implicated as a driver of Th17 responses in gnotobiotic mouse models. Nevertheless, more work is required to understand mechanistic links between disease processes and members of the gut microbiome.

In this study we have expanded on existing knowledge of the host transcriptome and mucosa-associated microbiome in PSC/UC across multiple relevant intestinal sites. Using RNA-seq we have explored transcriptomes across the terminal ileum, caecum and rectum, providing an in-depth assessment of tissue-defining gene expression signatures and how these signatures vary in patients with PSC/UC and UC relative to healthy controls. In addition to host transcriptomics, we have explored the mucosa-associated bacterial component of the microbiome using 16S ribosomal ribonucleic acid (rRNA) amplicon sequencing across the same intestinal sites, revealing modest associations in disease.

**Methods**

**Participant recruitment**

All study participants were recruited from the department of Gastroenterology, University of Oxford, and samples were collected as part of the Translational Gastroenterology Unit (TGU) biobank in the John Radcliffe Hospital in Oxford and part of the BRC Oxford IBD Cohort (REC 09/H1204/30) and BRC GI Illness Biobank (REC 16/YH/0247). Participants with a previous formal diagnosis of PSC/UC or UC according to clinical, radiological, endoscopic and histological criteria were identified from our databases and approached at outpatient clinic or endoscopy appointments to confirm consent and obtain intestinal samples. There were no specific inclusion/exclusion criteria. The sample size was based on what was achievable in the timeframe of the project.

For RNA-seq analysis, samples were collected from 13 control participants, 9 patients with PSC/UC and 10 patients with UC. Controls consisted of patients undergoing colonoscopy for chronic diarrhoea/abdominal pain, iron-deficiency anaemia, or follow-up after previous polyps, with findings of uninfamed intestine. Biopsies were taken from macroscopically and microscopically - as confirmed by an independent pathologist - normal colon and terminal ileum. For microbiome profiling, samples were collected from 16 control participants, 20 patients with PSC/UC and 18 patients with UC.

**Ethical considerations**

All patients provided written informed consent for participation in this study through the TGU Biobank (BRC Oxford IBD Cohort (University of Sheffield Research Ethics Committee (REC) 09/H1204/30) and BRC GI Illness Biobank (University of Sheffield REC 16/YH/0247).
Sampling intestinal sites through biopsy
Biopsies were obtained from ileum, caecum and rectum from patients with PSC/UC and patients with UC alone, undergoing colonoscopy for surveillance or disease activity and extension assessment, and non-inflammatory controls undergoing colonoscopy for colorectal cancer screening or chronic abdominal pain/diarrhoea with a macroscopically and microscopically normal colon. Endoscopic activity of PSC/UC and UC was defined by the Ulcerative Colitis Endoscopic Index of Severity (UCEIS) score2021 and microscopic inflammation by the Nancy score.22 Other clinical and demographic information, including current medications, were obtained from our clinical patient database.

Nucleic acid extraction
Biopsies were collected in RNAlater™ TissueProtect (Qiagen) and stored according to manufacturer guidelines. On the day of DNA/RNA extraction, samples were thawed at room temperature, RNAlater was removed and 350–700μl of RLT Buffer (Qiagen) with 10μl β-mercaptoethanol (Sigma) was added to each sample and samples were transferred to Soft Tissue Homogenizing CK14 Kit vials (Stretton Scientific). Samples were processed using Frewell’s 24 (Bertin Instruments) homogeniser (40s at 6500 RPM). 350μL of 70% ethanol was added to homogenised tissue lysates and lysates were transferred to QIAshredder column (Qiagen cat 79656) and subsequently transferred to the AllPrep DNA/RNA MicroKit (Qiagen). An on-column DNase digestion was performed for each sample using the RNase Free DNase Kit (Qiagen). Samples were eluted from columns in 30μL RNAse free; DNAse free H2O and RNA concentrations and 260/280 ratios were determined on a NanoDrop 1000 (Thermo Fisher).

RNA-seq
Selection of PolyA+ mRNA, generation of double stranded cDNA and library construction were performed using NEBNext Poly(A) mRNA Magnetic Isolation Module (E7490) and NEBNext Ultra II Directional RNA Library Prep Kit for Illumina (E7760L) with Illumina adapters and barcode tags (dual indexing, based on 23). The RNA-seq libraries were pooled and sequenced on an Illumina NovaSeq6000 as 150bp paired end reads. This generated an average of 71.7M (324.5K – 104.5M) paired reads per sample. We removed one sample from downstream analysis based on markedly low read counts (324.5K). The remaining samples had an average of 72.4M (27.5M – 104.5M) paired reads.

16S rRNA amplicon sequencing
For bacterial 16S rRNA gene amplicon sequencing, the variable V3 and V4 regions of the 16S rRNA gene were amplified by genomic DNA using the primers (standard IUPAC nucleotide nomenclature): Forward Primer = 5’TCGTCGGCAGCGTCAAGATGTATAGAAAACTACGGGNGGCWCCGAGTCGTCTCGGTCACGACGCTTCAGCGTGAAGAAAGAGACTTACGGTATCTAATCC
Reverse Primer = 5’GACAGTGATCGTCAAGAGACAGGA
The amplicons were then attached with indices and Illumina sequencing adapters using the Nextera XT index kit. The 16S amplicon libraries were pooled and sequenced in an Illumina MiSeq v3 flowcell as 300 paired end reads. Samples were removed from downstream analysis if they had fewer than 10000 paired reads (n = 19).

Processing RNA-seq data
Raw reads were assessed for quality using Fastqc (v0.11.7). Paired reads were aligned to the human genome (build hg38) using HISAT22425 (v2.1.0). Gene counts were produced using featureCounts26 (v1.6.0) against Ensembl genesets (Ensembl build 9127). Initial principal components analysis of transformed read counts (log2 counts per million) across tissues revealed three outlying patient samples. The patterns of gene expression were indicative of a sample swap and these samples were removed from further analysis. The resulting dataset therefore consisted of PSC/UC (n = 7), UC (n = 10) and healthy controls (n = 12).

Generating tissue-defining transcriptional signatures
In all analyses of RNA-seq data, genes that had ≥ 10 reads in ≥ 7 samples (minimum number of samples for any tissue-disease group) were taken forward. We used the likelihood ratio test (LRT) in DESeq228 to determine transcriptional differences among tissue locations comparing ileum, caecum and rectum samples from healthy individuals in a single test. Given that all participants contributed samples from each tissue location, the effect of individual variation was controlled for in the analysis (i.e. the DESeq2 design was specified as ~Individual + tissue). Genes were considered as significantly varying by tissue at an adjusted p-value < 0.05. Significantly varying genes were visualised in a heatmap and clustered using Manhattan distance and Ward.D clustering (pheatmap, R3.6.1). Genes were assigned to co-varying clusters by passing an hclust object (Manhattan distance and Ward.D clustering) to the dynamicTreeCut package29. Using method="hybrid" and cutHeight=30000 as parameters to dynamicTreeCut led to the identification of five clusters of genes.

Gene ontology pathway enrichment analysis
Pathway enrichment in tissue-defining gene clusters was performed using the hypergeometric test implemented in runGO.py from cgat-apps30. Genes from each cluster were output as the foreground datasets with all genes that had been assigned to a cluster used as the background set. The Gene Ontology (GO) biological pathway gene sets available from the Molecular Signatures Database (MSigDB, c5.bp.v6.1) was formatted using gmt2tsv.py to make it compatible with runGO.py. Gene sets were considered enriched at a Benjamini-Hochberg adjusted p-value < 0.05.

Differential expression across disease groups
We performed differential expression analysis in a pairwise manner (i.e. PSC/UC vs healthy, UC vs. healthy and PSC/UC vs. UC) in each tissue location separately, using DESeq2. Genes were considered significantly differentially expressed at an...
adjusted p-value < 0.05. Results from each of these comparisons were compared to each other using a combination of Venn diagrams (VennDiagram, R3.6.1) and scatterplots of log2(fold changes).

Enrichment of tissue-defining clusters among disease-associated gene sets
We used our previously defined tissue-defining gene clusters as input gene sets into runGO.py. Foreground gene sets were those genes that were called as differentially expressed in the pairwise differential expression analyses across diseases. The background gene set was any gene that had been assigned to a cluster. Clusters enriched in differential gene sets were considered significant at a Benjamini-Hochberg adjusted p-value < 0.05.

Analysis of gamma-glutamyl transferase in intestinal single-cell RNA-seq data
Single-cell transcriptomic data10 with accession number GSE125970 were downloaded from the gene expression omnibus (GEO). Processed data that represented unique molecular identifier (UMI) counts were used for the analyses (GSE125970_filtered_dataset.txt). Therefore annotated cell clusters that had been previously described10 were used to determine epithelial cell types that expressed GGT1. Any cell with at least one UMI annotated to GGT1 was considered GGT1 positive (GGT1+).

Cell deconvolution analysis
To predict changes in cell composition between groups we used CIBERSORTx31 for cell deconvolution. We utilised single cell RNA-seq data of immune cells from the human gut12 as the input signatures matrix to CIBERSORTx. This gave us the capacity to predict cellular fractions of 25 immune cell types (IgA plasma B cell, memory B cell, CD8 T cell, gamma-delta T cell, mast cell, innate lymphoid cell, macrophage, natural killer cell, follicular B cell, IgG plasma B cell, Central memory T cell, cycling B cell, Treg, LYN1 macrophage, Th1 cell, Th17 cell, conventional dendritic cell 1, conventional dendritic cell 2, cycling gamma-delta T cell, monocyte, activated CD4 T cell, T follicular helper cell, lymphoid dendritic cell, peripheral dendritic cell and cycling dendritic cell). Differential frequency of each cell type between either PSC/UC vs. healthy or UC vs. healthy was determined using a Wilcoxon rank sum test in R3.6.1. Significant differences were determined at a Benjamini-Hochberg adjusted p-value < 0.05.

Processing 16S rRNA amplicon sequencing
Raw reads were assessed for quality using Fastqc (v0.11.7). Given a minimal overlapping region between paired reads, we continued analysis using the first read of the pair (V3 variable region). Reads were processed using DADA233, implemented using a CGAT-core46 based pipeline (pipeline_dada2.py) available at https://github.com/OxfordCMS/OCMS_16S. First, we trimmed reads to 200bp and removed primer sequences (17bp). The resulting 183bp reads were aligned against the human genome (hg38) using bowtie257 (v2.3.4.1) and any aligned reads were filtered out. Alignment and filtering were performed using a CGAT-core based pipeline (pipeline_filter.py) available at GitHub. Filtered reads were then taken through the DADA2 workflow of learning error rates, sample inference and taxonomy assignment. Taxonomy was assigned using the RefSeq-RDP16S_v2_May2018 database. Downstream analysis was performed in amplicon sequence variant (ASV) counts summed at the level of genus.

Analysis of 16S rRNA amplicon sequencing controls
To aid interpretation of our 16S rRNA amplicon sequencing data we included a series of controls in the experiment. First, we utilised a mouse model (C57BL/6) with a defined microbial community with 12 members (Acetobacter muris, Akkermansia muciniphila YL44, Bacteroides caccae vivus YL148, Bifidobacterium animalis YL2, Blautia cocoides YL58, Clostridium ciastridiforme YL32, Clostridium innocuum 146, Enterococcus faecalis KB1, Flavonifractor plautii YL31, Lactobacillus reuteri 149, Maribaculum intestinale YL27 and Turicicmonas muris YL45)38. All experiments were conducted in accordance with the UK Scientific Procedures Act (1986) under a Project License (PPL) authorized by the UK Home Office (PS08F/A1F). Animals were housed in accredited animal facilities at the University of Oxford and provided with sterile water and food ad libitum, and environmental enrichment. This section of the manuscript is reported in line with the Animal Research: Reporting of In Vivo Experiments (ARRIVE) guidelines47. Germ free C57BL/6J mice were stably colonized with a defined consortium of 12 bacterial members of murine intestinal flora to generate a sDMDMm2 (MM12) colony that was maintained in sterile flexible film isolators.

We serially diluted (1:10 series) caecal contents from a single mouse from this colony (1-1:100000) reasoning that real contaminants would show up in the more dilute samples as has been previously described in single colony experiments48. All samples that were extracted were included in the analysis and no criteria were set for exclusion. The least dilute samples acted as a pseudo-positive control i.e. we could determine whether we were able to sequence members of the community but not accuracy in relative abundance estimations (we did not know this beforehand). Any ASVs that were assigned to a genus that was not part of this experimental community was flagged as a potential contaminant. In addition to these controls we also sequenced a negative control sample. This was extracted DNA from an endoscopic brush that had been passed through the endoscope during colonoscopy, once the caecum was reached but had not been used to brush the mucosa. Any ASV found in this endoscopy control sample was added to the list of potential contaminants for future cross-referencing.

Exploratory analysis of 16S rRNA amplicon data
We performed alpha diversity analysis using the phyloseq45 package (R3.6.1) and significance of disease group differences was determined using the Kruskall-Wallis test (Kruskal.test in R3.6.1). We explored beta-diversity associations with variables of interest (tissue location, experimental group) and covariates (sequencing batch and DNA extraction kit) using principal components analysis (PCA) of log10-transformed genus relative abundance estimates (prcomp in R3.6.1). We defined significant
associations as a p-value < 0.05 in a univariate Adonis test (vegan package in R3.6.1) run for each variable separately. Given that we observed a significant effect (p < 0.05) of sequencing batch, we included this as a covariate in Adonis testing of both tissue location and disease. To determine the relationship between inter-individual and inter-tissue variability we performed dissimilarity analysis (Bray-Curtis dissimilarity implemented in phyloseq) both within-tissue sites (i.e. between individuals) and within individuals (i.e. between tissue-locations). Significant differences were determined using the Wilcoxon Rank Sum Test (wilcox.test in R3.6.1) using within-individual dissimilarity as the reference for comparisons.

Genus differential abundance testing between disease groups
We used DESeq2\textsuperscript{23} to assess differential genus abundance between disease groups. This was performed in a pairwise manner between each disease group, in each tissue separately. Genera were deemed significantly differentially abundant at an adjusted p-value < 0.05.

Host-microbiome correlation analysis
The method we employed for assessing host-microbiome correlations is outlined in Extended data Figure S3\textsuperscript{37}.

To remove the effect of disease on both host transcription and genus abundance, and thus the potential for spurious correlations, we first removed effects of disease status using Combat-seq\textsuperscript{40}. Processed count matrices were taken forward for downstream analysis. Host gene expression counts were normalised using the log2(counts per million) and these values were z-score scaled (across samples) to eliminate effects of expression level when defining co-expressed genes. Co-expression modules for each tissue location were defined using WGCNA\textsuperscript{45} with default parameters. Each module in each tissue was assessed for enrichment of GO biological pathways using GOseq\textsuperscript{42}. If significant enrichments were identified (padj < 0.05) then the module was annotated with the top pathway i.e. smallest adjusted p-value. Centred log-ratio transform using ALDEx2\textsuperscript{43} was used on genus counts and these were used for correlation analysis. Pearson product-moment correlations were used to correlate all module eigengenes (i.e. the first principal component of standardised expression levels for each module) and non-zero genus abundances using the cor.test function from \texttt{psych}\textsuperscript{44}. Correlations were retained if they reached an adjusted p-value < 0.05. As module eigengenes provide a representative expression summary for each module, we used significant correlations to conclude broad functional associations. Resulting networks were visualised using igraph v1.2.6\textsuperscript{45}.

Analysis code can be found as extended data\textsuperscript{16}.

Results
Tissue-defining transcriptional signatures across intestinal tissue sites
The number of participants and sample characteristics are provided in Table 1.

There were no significant differences between groups for demographic variables (age and sex, Extended data Figure S1A\textsuperscript{37}). Significant differences were found for medication use (Extended data Figure S1B\textsuperscript{37}). These included significant differences in the proportion of individuals receiving ursodeoxycholic acid (URSO) between groups (PSC/UC (100%), UC (0%), healthy (0%), p = 6.4×10\textsuperscript{-7}) and aminosalicylates (PSC/UC (100%), UC (80%), healthy (0%), p = 1.3×10\textsuperscript{-4}). The majority of patients with either PSC/UC or UC did not have active colitis (Extended data Figure S1C\textsuperscript{37}) and there were no significant differences between patients with PSC/UC and patients with UC for neither the Nancy index (p = 0.26) nor the UCEIS (p = 1). While higher in a subset of patients with PSC/UC, liver function measures (Bilirubin, ALT and ALP) were not significantly different between groups (Extended data Figure S1D\textsuperscript{37}).

Given there is known functional variation between intestinal locations, first we aimed to assess how global gene expression profiles vary by both tissue location and disease status. We generated paired RNA-seq data sets from biopsies taken from the ileum, caecum and rectum of patients with PSC/UC (n = 8), UC (n = 10) and healthy individuals (n = 12) (Figure 1A). Principal components analysis (PCA) revealed tissue location as a major source of gene expression variation. Ileal samples separated from both caecal and rectal samples along principal component (PC) 1 (39%), with caecal and rectal samples displaying similar yet distinct transcriptional profiles (PC3, 9%) (Figure 1B). To examine in more detail genes and gene sets that define different tissue locations, we performed differential expression analysis and identified 13,071 genes (60% of all genes analysed) as differentially expressed (Extended data Table S1\textsuperscript{37}) between tissues in healthy individuals. These genes could be further partitioned into five clusters that reflected genes that were highest in the ileum (cluster 1), high in the caecum and rectum relative to the ileum (cluster 2), highest in the rectum (cluster 3), a mixed cluster that contained genes that were either high in the caecum or high in the ileum, and genes that were high in a subset of healthy ileal samples (cluster 5) (Figure 1C and Extended data Table S2\textsuperscript{37}). Ileum-defining genes (cluster 1, 3218 genes) were enriched for gene ontology biological pathways (GO:BP) associated with lipid and drug metabolism (Figure 1D). Caecum/rectum-defining genes (cluster 2, 3082 genes) were enriched for GO:BP involved in cell cycle and amino acid activation (Figure 1E). Rectum-defining genes (cluster3, 2853 genes) were significantly enriched for GO:BP involved in neurogenesis and neurotransmitter release (Figure 1F). Cluster 4 genes (2271 genes) were enriched for GO:BP linked with adaptive immune responses (Figure 1G). Finally cluster 5 (1647 genes) were not enriched for any GO:BP. Enrichment results are available as Extended data tables (Extended data Tables S3–6\textsuperscript{37}).

Having mapped tissue-defining pathways, we reasoned that these functional differences would at least in part be driven by tissue-dependent transcription factor profiles. We explored this by assessing enrichment of transcription factor binding motifs amongst genes in each cluster. Each cluster was associated with specific transcription factors (Extended data
Table 1. **Patient and sample characteristics.** Quantitative variables are expressed as median (standard deviation) and categorical variables as number of samples. Numbers are broken down by tissue site as not all patients had a sample for every tissue site.

### RNA-seq samples

<table>
<thead>
<tr>
<th></th>
<th>Control</th>
<th>PSC/UC</th>
<th>UC</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Ileum</td>
<td>Caecum</td>
<td>Rectum</td>
</tr>
<tr>
<td><strong>Demographics</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Age (Years)</td>
<td>46 (14.18)</td>
<td>42 (15.17)</td>
<td>49 (17.00)</td>
</tr>
<tr>
<td>Sex (F/M)</td>
<td>2/9</td>
<td>1/8</td>
<td>3/9</td>
</tr>
<tr>
<td><strong>Inflammation</strong></td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>UCEIS</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td><strong>Medication</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>URSO</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Vedolizumab</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>ASA</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Azathioprine</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Steroids</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td><strong>Liver function</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bilirubin (nMol/L)</td>
<td>12 (3.56)</td>
<td>9.5 (3.56)</td>
<td>12 (3.56)</td>
</tr>
<tr>
<td>ALP (IU/L)</td>
<td>86 (26.87)</td>
<td>86 (30.72)</td>
<td>86 (26.87)</td>
</tr>
<tr>
<td>ALT (IU/L)</td>
<td>22 (11.92)</td>
<td>28 (10.90)</td>
<td>22 (11.92)</td>
</tr>
</tbody>
</table>

### 16S amplicon sequencing samples

<table>
<thead>
<tr>
<th></th>
<th>Control</th>
<th>PSC/UC</th>
<th>UC</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Ileum</td>
<td>Caecum</td>
<td>Rectum</td>
</tr>
<tr>
<td><strong>Demographics</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Age (Years)</td>
<td>45 (16.29)</td>
<td>45 (16.90)</td>
<td>57 (19.37)</td>
</tr>
<tr>
<td>Sex (F/M)</td>
<td>5/9</td>
<td>5/9</td>
<td>4/9</td>
</tr>
<tr>
<td><strong>Inflammation</strong></td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>UCEIS</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td><strong>Medication</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>URSO</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Vedolizumab</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>ASA</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Azathioprine</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Steroids</td>
<td>0</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td><strong>Liver function</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bilirubin (nMol/L)</td>
<td>10 (4.24)</td>
<td>7 (4.21)</td>
<td>8.5 (3.66)</td>
</tr>
<tr>
<td>ALP (IU/L)</td>
<td>82 (23.03)</td>
<td>83 (25.48)</td>
<td>81.5 (21.61)</td>
</tr>
<tr>
<td>ALT (IU/L)</td>
<td>23 (8.91)</td>
<td>22 (9.51)</td>
<td>21.5 (9.51)</td>
</tr>
</tbody>
</table>

UCEIS = Ulcerative Colitis Endoscopic Index of Severity. Nancy = Nancy histologic index. URSO = Ursodeoxycholic acid. ASA = 5-aminosalicylates. ALP = Alkaline Phosphatase. ALT = Alanine aminotransferase.
Figure 1. Tissue-defining transcriptional signatures. (A) Experimental design and number of samples obtained for ribonucleic acid (RNA)-seq. (B) Principal components analysis (PCA) showing the relationship between all samples (PCA performed on log2(counts per million) with the `prcomp` function in R3.6.1). (C) Heatmap (Manhattan distance, Ward.D2 clustering) showing genes called as significantly differentially abundant across tissues in healthy individuals. Clusters defined using the `dynamicTreeCut` method are displayed as row annotations. (D–G) Top 10 gene ontology biological pathways (GO:BP) that are significantly enriched (false discovery rate (FDR) < 0.05) amongst genes in each tissue-defining cluster. GO categories are coloured (filled circles) based on manual annotation of GO:BP. (H–K) Top 10 transcription motif gene sets (MsigDB, C3, regulatory targets) that are significantly enriched (FDR < 0.05) amongst genes in each tissue-defining cluster.
Tables S7–10). Cluster 1 (ileum-defining) was associated with genes that are predicted to be regulated by hepatic nuclear factors (HNF) (Figure 1H), cluster 2 (caecum/rectum-defining) with E2F transcription factors involved in the cell cycle (Figure 1I), cluster 3 (rectum-defining) genes with serum response factor (SRF) transcription factors (Figure 1J) and cluster 4 (mixed ileum and caecum) genes with erythroblast transformation specific (ETS) family transcription factors including PU.1 (Figure 1K).

Consistent with the transcription factor binding predictions, we observed significant increased expression of hepatic nuclear factor 1A (HNF1A) in the ileum (cluster 1), along with genes predicted to be regulated by HNF and involved in lipid metabolism (G6PC, SOAT2 and UGT1A1, Figure 2A). We also observed significantly increased expression of the E2F transcription factor, E2F1, that regulates key cell-cycle genes CDK1, MCM3 and MSH2 in caecal/rectal samples (cluster 2, Figure 2B). While SRF itself was not significantly up-regulated in rectal samples, SRF-regulated genes involved in neurogenesis including HOXD10, INSM1, NRF2 and DIXDC1 were associated with the rectal transcriptome (cluster 3, Figure 2C). The ETS family member PU.1 is known to regulate myeloid cell activation. We found predicted targets of ETS/PU.1 to have increased expression in caecal samples. These up-regulated genes included classical markers of myeloid cell activation (CD86, MEF2C, FCER1G and TLR4, Figure 2D). These data suggested an increased antigen-driven cellular activation of myeloid cells in healthy caecum relative to ileum and rectum. Indeed, in addition to a proposed myeloid cell activation, we observed an increased expression of immunoglobulin genes (Figure 2E) in caecal samples relative to ileum and rectum, supporting an increased proportion of activated B cells in this tissue.

Together, these data support tissue-dependent functions that reflect HNF-regulated lipid metabolism in the ileum, SRF-dependent neuronal function in the rectum and antigen-dependent lymphocyte activation in the caecum.

Tissue-dependent changes in gene expression in PSC/UC and UC

Having established tissue-defining gene expression profiles across healthy intestinal sites, we next aimed to explore gene expression alterations in disease and how they map onto tissue signatures.

We determined gene expression differences between healthy individuals and disease (PSC/UC and UC) at each tissue location (Extended data Tables S11–16). A summary of the number of changes in each comparison is shown in Table 2 (false discovery rate (FDR) < 0.05). The total number of differentially expressed genes varied between each comparison (Figure 3A–C), with differences being largely tissue dependent (Figure 3D). While there was little overlap in genes differentially regulated in each disease relative to healthy individuals (Figure 3E–G), effect sizes (i.e. log2(fold change)) were well correlated, particularly for effects observed in the ileum (Figure 3H) and to a lesser extent in the caecum (Figure 3I). In addition, while correlated, effect sizes in the rectum were higher in UC when compared with healthy than PSC/UC compared to healthy (Figure 3J). This is consistent with clinical features of these diseases, with rectal involvement predominantly being a feature of UC as opposed to PSC/UC.

Given a high degree of tissue dependency on disease-associated gene expression, we wanted to explore whether this was due to inherent differences in tissue gene expression. To explore this, we utilised our previously identified tissue-defining gene clusters to define overrepresented tissue-dependent processes in disease-associated gene lists. In the ileum, disease-associated transcriptional changes were not associated with ileum-defining transcriptional signatures in general. Indeed, changes in both PSC/UC and UC were enriched for genes that typically define colonic tissue sites (Clusters 2 and 4, Figure 4A and B). For example, the expression of ribosomal proteins that define a caecum/rectum profile (cluster 2) were increased in expression in both PSC/UC and UC when compared to healthy controls (Figure 4E). These changes may be partially explained by an increase in proliferating lymphocytes in disease, as we also observed an increase in expression of the lymphocyte-specific helicase, HELLS (Figure 4E), particularly in PSC/UC. In the caecum, PSC/UC-associated changes were enriched for genes belonging to cluster 4 (Figure 4C). These changes are associated with immune signatures that appear to define T cell alterations in PSC/UC, and to a lesser extent UC. Such changes are exemplified by reduced expression of key T cell markers including CD2, CD3D, CD3E and CD3G (Figure 4E). These general changes may reflect alterations in specific T cell populations as both IL7R and FOXP3 (Figure 4E) were reduced in PSC/UC. To further explore whether there was a reduction in specific immune cell populations we performed cell deconvolution analysis using CIBERSORTx. Prediction of cellular fractions from 25 human gut cell types revealed no statistically significant differences in cellular composition between disease groups in any of the three tissue types. Further work using single cell measurements with either flow cytometry or single cell RNA-seq would be required to accurately deconvolve this signal.

In UC, changes in the rectum relative to healthy individuals were enriched for genes in cluster 4 (Figure 4D) and include ALG1L, RFXANK, LITD1 and ZMAT5 (Figure 4N–Q). These changes represent a relatively broad altered functional profile that includes increased expression of genes involved in glycosylation (ALG1L), self-renewal (LITD1) and MHC class II processing (RFXANK).

Gamma-glutamyl transferase 1 (GGT1>) is specifically reduced in PSC/UC

We directly compared PSC/UC to UC to identify genes that were specifically altered in PSC/UC. In total, three, sixteen, and nine genes were differentially expressed in ileum, caecum and rectum, respectively (Extended data Tables S17–19). Of these, GGT1 was significantly reduced in both the ileum and caecum in patients with PSC/UC relative to both healthy individuals and patients with UC (Figure 5A). In contrast to the ileum and caecum, we did not observe a significant reduction in GGT1 expression in the colon of patients with PSC/UC.
Figure 2. Tissue-defining gene expression signatures support distinct transcription factor profiles. (A) Genes in cluster 1 are enriched for genes with HNF transcription factor motifs in their promoters. Genes plotted represent genes that were found in gene ontology biological pathways (GO:BP) annotated as being involved in lipid metabolism and also predicted to be regulated by HNF transcription factors (MSigDB, C3 regulatory targets gene sets). (B) Genes in cluster 2 are enriched for genes with E2F transcription factor motifs in their promoters. Genes plotted represent genes that were found in GO:BP annotated as being involved in cell cycle progression and also predicted to be regulated by E2F transcription factors (MSigDB, C3 regulatory targets gene sets). (C) Genes in cluster 3 are enriched for genes with SRF transcription factor motifs in their promoters. Genes plotted represent genes that were found in GO:BP annotated as being involved in neurogenesis and also predicted to be regulated by SRF transcription factors (MSigDB, C3 regulatory targets gene sets). (D) Genes in cluster 4 are enriched for genes with ETS/PU.1 transcription factor motifs in their promoters. Genes plotted represent genes that were found in GO:BP annotated as being involved in adaptive immune responses and also predicted to be regulated by PU.1 transcription factors (MSigDB, C3 regulatory targets gene sets). (E) Genes in cluster 4 that also contribute to a signal of adaptive immune responses are multiple immunoglobulin genes.
Table 2. Number of genes called as significantly differentially expressed in each comparison (false discovery rate (FDR) < 0.05).

<table>
<thead>
<tr>
<th></th>
<th>Ileum (up)</th>
<th>Ileum (down)</th>
<th>Caecum (up)</th>
<th>Caecum (down)</th>
<th>Rectum (up)</th>
<th>Rectum (down)</th>
</tr>
</thead>
<tbody>
<tr>
<td>PSC/UC vs. healthy</td>
<td>151</td>
<td>45</td>
<td>46</td>
<td>174</td>
<td>6</td>
<td>0</td>
</tr>
<tr>
<td>UC vs. healthy</td>
<td>427</td>
<td>302</td>
<td>1</td>
<td>2</td>
<td>30</td>
<td>7</td>
</tr>
</tbody>
</table>

Figure 3. Differentially expressed genes in the ileum, caecum and rectum between disease and healthy. Heatmaps (Manhattan distance with Ward.D2 clustering) of log2(CPM) values for the union of differentially expressed genes (DESeq2 adjusted p-value < 0.05) in disease (either PSC/UC or UC) vs. healthy in ileum (A), caecum (B) and rectum (C) are shown. (D) Overlap between tissue sites for the union (as described) of disease-associated genes. Venn diagrams for the overlap of genes in PSC/UC vs healthy and UC vs. healthy in the ileum (E), caecum (F) and rectum (G) are displayed. The overall correlation of effect sizes (log2(fold change)) between PSC/UC vs. healthy and UC vs. healthy in the ileum (H), caecum (I) and rectum (J) are shown. Correlations and significance calculated using the Pearson correlation coefficient implemented in the cor.test function in R3.6.1. PSC/UC=primary sclerosing cholangitis-associated ulcerative colitis.

in an independent, publicly available data set\(^8\) suggesting that this alteration is specific to the small intestine and proximal colon.

Increased serum gamma-glutamyl transferase (GGT) is a diagnostic marker of alcoholic liver disease\(^{50}\) and cholestasis in children\(^5\). Functionally, it participates in the glutathione salvage pathway, enabling glutathione-dependent reduction of reactive oxygen species in the cell. In the kidneys, reduced GGT1 expression is linked to mitochondrial dysfunction and metabolic re-wiring in a relatively rare renal cancer - Chromophobe renal cell carcinoma (ChRCC)\(^52\). We therefore reasoned that this
enzyme may be an important modulator of colon cancer risk in PSC/UC and as such would be expressed in intestinal epithelium. To examine this, we leveraged publicly available single cell RNA-seq data generated from the epithelial compartment of human ileum, colon and rectum. As described in the original publication, the ileal epithelium differed substantially from the colonic and rectal epithelium in cellular composition, particularly in terms of an increased proportion of enterocytes (Figure 5B). Cells expressing GGT1 (GGT1+ >= read) were enriched in enterocytes (Figure 5C), a feature that contributed to an observed increase in proportion of GGT1+ cells in the ileum when compared to the colon and rectum (Figure 5D). These data support expression of GGT1 in the enterocyte compartment of the ileal and colonic epithelium. A defining feature of PSC/UC compared with UC is liver disease and cholestasis. Given the strong functional links between the liver and the intestine, particularly with respect to bile acid secretion, our results suggest that further work should aim to identify links between altered liver function and downstream effects on intestinal GGT1 expression in patients with PSC/UC.

Mucosa-associated bacterial associations with disease
The intestinal microbiota performs essential homeostatic roles and is an important mediator of the liver-gut axis that includes

---

**Figure 4. Enrichment of tissue-defining clusters amongst disease-associated genes.** (A–B) Genes identified as significantly differentially expressed between PSC/UC vs. healthy or UC vs. healthy in the ileum were tested for enrichment for each previously described tissue-defining cluster. Asterisks denote significant enrichment of a given cluster at a Benjamini-Hochberg adjusted p-value < 0.05 (hypergeometric test). (C) Genes identified as significantly differentially expressed between PSC/UC vs. healthy in the caecum were tested for enrichment for each previously described tissue-defining cluster. No significant enrichment was found for genes differentially expressed in UC vs. healthy caecum and as such the plot is not shown. (D) Genes identified as significantly differentially expressed between UC vs. healthy rectum were tested for enrichment for each previously described tissue-defining cluster. No significant enrichment was found for genes differentially expressed in PSC/UC vs. healthy rectum and as such the plot is not shown. (E) Heatmap visualisation of significantly differentially expressed genes between PSC/UC vs. healthy or UC vs. healthy that belong to significantly enriched clusters at each tissue site. Clusters are provided as left annotations and the tissue site where differential expression was observed is also annotated. Heatmaps were produced using the ComplexHeatmap package in R 3.6.1. As numerous ribosomal proteins were observed to be differentially expressed in the ileum, these are also annotated. PSC/UC=primary sclerosing cholangitis-associated ulcerative colitis.
the deconjugation of liver-derived bilirubin and bile acids. Given functional differences and microbiome composition differences between intestinal sites, we aimed to assess bacterial associations in PSC/UC of the mucosa-associated microbiota across the ileum, caecum and rectum. We performed 16s rRNA amplicon sequencing (Figure 6A) of biopsy samples from the same three tissue sites as described (ileum, caecum and rectum) in healthy individuals, patients with PSC/UC and patients with UC.

Samples were removed from the analysis if they had fewer than 10000 reads. The mean number of reads remaining after filtering for host-associated reads and processing with dada2 was 95617 (10542–543905). On average, we could assign 99% of sequence reads to five phyla, with all samples dominated by three major phyla typical of intestinal microbiomes (Firmicutes, Bacteroidetes and Proteobacteria, Figure 6B–D). Using amplicon sequence variant (ASV) abundances, we observed no significant association between disease status and Shannon diversity in the ileum (F(2,45) = 1.90, p = 0.16), caecum (F(2,44) = 1.88, p = 0.17, or rectum (F(2,41) = 0.16, p = 0.85)). Further downstream analyses were performed at the genus level i.e. ASV counts summed at the level of genus (total number of genera analysed was 128). In contrast to host transcriptomes, overall microbiome composition was not associated with tissue location (Figure 6E). However, there was a significant clustering by disease status (Figure 6F) and sequencing batch (Figure 6G). Nevertheless, the variation in the bacterial component of the microbiome was driven by inter-individual differences. Indeed, samples from the same individual (across tissue locations) were more similar to each other than samples from the same tissue location (between individuals) (Figure 6H).

Following a similar approach as previously described for assessing potential reagent contamination, using serially diluted (1:1 – 1:1000,000) mouse caecal content samples from mice with a defined 12-member community we observed little evidence for reagent contamination. Indeed, between 96.3% and 100% of sequenced reads were assigned to genera within this community (Extended data Figure S2). All expected genera were recovered from this community, supporting a reasonable extraction and sequencing approach in this study. Nevertheless, using this approach, as well as sequencing data from an endoscope control (negative control), we identified 135 ASVs that we considered as potential contaminants (Extended data Table S20). While these sequences were retained for downstream analysis, we used them as a reference when reviewing significant disease associations.
Figure 6. Mucosa-associated bacterial associations in PSC/UC and UC. (A) Description of the number of samples collected for 16S ribosomal ribonucleic acid (rRNA) amplicon sequencing. High-level phylum-level classification and abundances are provided for the ileum (B), caecum (C) and rectum (D). Principal components analysis (PCA) plots of log10(relative abundance + 1E-6) at the genus level showing clustering by tissue (E), disease group (F) and sequencing batch (G) are shown. $R^2$ and p-values were computed using the `adonis` test in the `vegan` package in R3.6.1. For the assessment of both tissue and disease associations, sequencing batch was included as a covariate. (H) Violin plot displaying inter- and within-individual dissimilarity measured by Bray-Curtis dissimilarity on relative abundance estimates. Bray-Curtis dissimilarity was computed using the `Phyloseq` package in R3.6.1. (I) Differentially abundant (DESeq2 adjusted p-value < 0.05) genera in UC vs. healthy ileum. (J) Differentially abundant (DESeq2 adjusted p-value < 0.05) genera in UC vs. healthy caecum. (K) Differentially abundant (DESeq2 adjusted p-value < 0.05) genera in PSC/UC vs. healthy rectum. Asterisks denote DESeq2 adjusted p-value < 0.05. (L) Detection of likely contaminants in control samples. CONTROL = Endoscope control, dil = dilution from dil0 (caecal contents 1:1) – dil6 (caecal contents 1:1,000,000). (M) Heatmap showing the distribution of potential contaminants amongst caecal samples. Row-scaled relative abundances are shown and metadata are labelled. The heatmap was produced using `pheatmap` with rows and columns clustered using Manhattan distance and Ward.D clustering. PSC/UC=primary sclerosing cholangitis-associated ulcerative colitis.
p < 0.05, Brucella, Herbaspirillum and Stentrophomonas, Figure 6I) and four in the caecum (Brucella, Herbaspi-
lillum, Stentrophomonas and Pseudoxanthomonas, Figure 6I), three of which overlapped between tissue sites (Brucella, Herbaspi-
lillum and Pseudoxanthomonas). One genus was found to be differentially abundant in the rectum (Parabacteroides, Figure 6K) between patients with PSC/UC and healthy controls (Extended data Tables S21–26). Given the relatively low prevalence and low abundance of Brucella, Herbaspi-
lillum, Stentrophomonas and Pseudoxanthomonas we examined whether there was evidence that these were contaminants. Her-
aspiillum, Stenotrophomonas and Pseudoxanthomonas have been reported as reagent contamination artefacts in previous studies37. Using our control data, we were able to identify Herbaspiillum and Pseudoxanthomonas as likely contaminants. Indeed, these genera were identified in the most dilute sam-
pl.es of our known microbial community controls (Figure 6L). We did not observe Brucella nor Stenotrophomonas in our con-
trol data sets. However, the observation that their abundance co-varyed with likely contaminants, as well as them only appear-
ing in one of the two sequencing batches (Figure 6M), suggests that they are part of the same contaminant profile.

Together these data support high inter-individual variability in mucosa-associated bacterial composition and little evidence for tissue- or disease-associated bacterial genera. Nevertheless, there is a decreased abundance of Parabacteroides in the rectum of patients with PSC/UC.

Exploring host-bacteria associations across tissue sites

Next, we sought to understand the relationship between host gene expression and bacterial relative abundances at each tissue site. To assess this, we correlated bacteria relative abundances with host gene expression modules derived from weighted gene co-expression network analysis (WGCNA). We identified 22, 28 and 23 host modules in the ileum, caecum and rec-
tum, respectively. Of these, 9 (40%), 20 (71%) and 16 (70%) could be annotated with at least one significantly enriched (padj < 0.05) GO:BP in the ileum, caecum and rectum, respectively.

Correlation analysis between all module eigengenes and cen-
tred log-ratio (CLR)-transformed genus counts revealed an influence of high numbers of zero values in genus abundances in driving significant host-bacteria correlations. For example, we observed a statistically significant (Benjamini-Hochberg adjusted p < 0.05) correlation between Corynebacterium abundance and a host module eigengene, ME25 (Extended data Figure S4A). However, due to the large number of zero counts in these data (Extended data Figure S4B), it was evident that this significant correlation was being driven by CLR-transformed zero counts (Extended data Figure S4C) leading to the conclu-
sion that this was a spurious correlation. To counter this effect, we re-ran the correlation analysis using only non-zero data. These analyses revealed little evidence for strong host-bacterial associations. Indeed, in all tissue sites, statistically signifi-
cant (Benjamini-Hochberg adjusted p < 0.05) correlations were restricted to intra-host or intra-bacterial correlations (Figure 7A–C). Our data suggest that host gene expression is not strongly associated with genus abundance.

Discussion

The human intestine is a multi-functioning organ that performs tasks from nutrient absorption and metabolism to protection against invading pathogens. In this study we have expanded on existing knowledge regarding the distribution of functions across different sites in the human intestine. In healthy individuals we observed increased expression of genes involved in lipid and small molecule metabolism in the ileum relative to both the caecum and rectum. These results were expected given previous observations of increased expression of genes involved in lipid assimilation in small intestinal epithelial cells relative to the colon and rectum16. Indeed, and while not discussed in detail here, we observed similar increased expression of APOAI and APOM in the ileum, as described previously16. We identified HNF transcription factors as key markers of ileal transcriptomic signatures. In addition to enrichment of HNF motifs among genes more highly expressed in the ileum relative to the caecum and rectum, we identified increased expression of HNF1α in the ileum. HNF1α is a key transcription factor in multiple metabolic processes across multiple cell types, with mutations in HNF1α causing maturity-onset diabetes of the young (MODY)15. Our data corroborate previous experiments in ani-
mal models suggesting that in addition to its role in glucose metabolism it has a prominent role in lipid metabolism16.

Both the caecum and the rectum were characterised by increased expression of genes related to cell cycle progression. Indeed, we observed an increased expression of the transcription factor E2F1 in the caecum and rectum relative to the ileum. E2F1 promotes cell cycle progression into S phase17 through its regu-
lation of replication-promoting target genes such as CDK718,19, a gene that we also observed to have increased expression in both the caecum and rectum relative to the ileum. These data support an increased frequency of proliferative cells in the colon relative to the small intestine. While we cannot determine the specific cell types that are proliferating from our data, it is likely that increased frequencies of proliferative stem, transit-
amplifying, and progenitor epithelial cells20 in the colon relative to the ileum16 contribute at least in part to our observations in bulk tissue samples.

We were able to further identify a cluster (cluster 4) of genes that represented transcription of immune processes – particularly adaptive immunity. While this cluster was somewhat mixed in terms of tissue distribution, we were able to identify transcriptional signatures of immune activation that were predominantly associated with increased transcription in the caecum relative to the ileum and rectum. These genes included CD86, MEF2C, FCER1G and TLR4. CD86 is expressed on the cell surface of antigen presenting cells (APCs) including dendritic cells, macro-
phages and B-cells, and participates in the activation of adaptive immunity through T-cell activation15. The increased expression of this marker in the caecum supports antigen-induced
activation at this tissue site\textsuperscript{62}. Further support for increased antigen-dependent activation of APCs in the caecum comes from observations of increased expression of \textit{FCER1G} and \textit{TLR4}, genes known to be involved in mediating innate control of adaptive immunity\textsuperscript{63}. The interactions of APCs with microbial antigens in the caecum could also explain increased expression of immunoglobulin genes in the caecum, which we propose represent an enrichment of activated B cells in the caecum relative to the ileum and rectum.

Together with the central nervous system (CNS), the enteric nervous system (ENS) is essential for digestion through controlling peristalsis and movement of fluids across the mucosa\textsuperscript{64}. We identified an enrichment of genes involved in neurogenesis that are more highly expressed in the rectum relative to both the ileum and caecum. Control of these genes was, at least in part, explained by an enrichment of serum response factor (SRF) motifs in the promoter regions of these genes. It is difficult to fully interpret these results as although SRF is a known neuroprotective factor in cortical neurons\textsuperscript{65}, our data do not support a general increase in neurons in the rectum relative to the ileum and caecum. For example, we do not see elevated expression of the neuronal marker \textit{ELAVL4} in the rectum. In contrast, we did observe an increased expression of the enteric glial cell (EGC) marker, \textit{GFAP}\textsuperscript{66}, in the rectum relative to both the ileum and caecum. Along with increased expression of the neuroendocrine monoamine transporter gene (\textit{SLC18A1}) and multiple genes involved in synapse formation (e.g., \textit{SNAP25}, \textit{Synaptogagmins}, Extended data Table 1), we would hypothesise that there is increased ENS-EGC crosstalk in the rectum relative to the ileum and caecum. Further work is required to determine if this is the case and what the functional consequences of such differences are.

One of the major aims of this study was to identify transcriptional profiles that vary across disease states and whether these are tissue-dependent. Differential expression analysis revealed multiple transcriptional differences between healthy individuals and individuals with either PSC/UC or UC across all three tissue sites. The majority of changes were identified in the
ileum, with a substantial overlap between PSC/UC and UC. This is a noteworthy observation as patients with UC rarely present with ileitis although it can be detected endoscopically in some cases

In the caecum, patterns of gene expression were consistent with clinical differences between PSC/UC and UC. Although effect sizes between PSC/UC vs. healthy and UC vs. healthy were fairly well correlated, we observed an expected increased effect size in PSC/UC caecum relative to UC caecum. This is consistent with more proximal involvement in PSC/UC relative to UC. In contrast, in the rectum, increased effect sizes and thus differential features were observed in UC relative to PSC/UC. These data are consistent with increased rectal involvement in UC relative to PSC/UC. It was noteworthy that differences in transcription between either PSC/UC or UC vs. healthy were not conserved across tissue sites. Initially we hypothesised that this reflected underlying differences between tissues that we had described previously. However, we found that changes in the ileum were not significantly enriched in genes that define the ileum relative to the caecum and rectum. Indeed, genes that are more highly expressed in either PSC/UC or UC relative to healthy controls were enriched for genes that are in cluster 2 of tissue-defining features – proliferation-related genes that are more highly expressed in the healthy caecum and rectum relative to the ileum. These changes consisted of multiple genes involved in protein synthesis, including multiple ribosomal proteins. We believe that these changes represent an increased in the frequency of proliferative immune cells in the ileum in patients with PSC/UC and UC. The higher expression of HELL5, a lymphocyte specific helicase, supports an increase in proliferating lymphocytes in the ileum, particularly in patients with PSC/UC. The increases in ribosomal proteins that we see also support cell growth/proliferation of immune cells. This is evident from cell-type specific expression patterns of a number of ribosomal proteins including RPL36, where expression is highest in T and B lymphocytes. Notably, we observed an increased expression of RPL5, a gene whose protein product is associated with P53-dependent and independent tumour suppressor function. While the role of RPL5 in reducing proliferation is incongruous with our data, recent evidence supports a role for P53 in regulating DNA polymerase restart during DNA replication, which may or may not involve RPL5.

We identified a clear T cell expression phenotype in the caecum of patients with PSC/UC relative to healthy controls. These changes may represent a reduced frequency of T cells as classical T cell markers such as CD2, CD3E, CD3F and CD3G were reduced in expression. Reduction in additional markers of T cell subsets including IL7R and FOXP3 also point to a reduction in specific functional subsets, particularly in PSC/UC. However, we did not observe a significant association with any predicted cell fraction using cell deconvolution analysis. This may reflect complexity in such analyses, that cannot successfully decompose individual cellular expression from expression governed by cellular composition. Further work would be required to capture single cell measurements of gene or protein expression to confirm our data.

Rectal involvement is more frequent in UC compared to PSC/UC. This clinical feature was borne out in the results of our differential expression analysis whereby a greater number of differential features with larger effect sizes were identified in the UC vs. healthy comparison relative to the PSC/UC vs. healthy comparison. Differentially expressed genes in the rectum of patients with UC represent a diverse set of functions. These include glycosylation, self-renewal and MHC class II processing. It is not clear how these genes might contribute to pathology in UC although it is noteworthy that increased expression of ALG1I in the rectum has been reported previously. Interestingly LITD1 is associated with self-renewal and the potential to promote cancer. Nevertheless, a recent report suggested that increased LITD1 expression is associated with increased survival in colorectal cancer. Further work is required to determine any functional impact of increased LITD1 in the rectum of patients with UC.

A prominent clinical difference between PSC/UC and UC is a significant increased risk of colorectal cancer in patients with PSC/UC relative to UC. Pathways underlying this difference remain unclear, although potential mechanisms could involve altered tumour-promoting immune phenotypes, DNA damage pathways and/or changes in the abundance of carcinogenic microbes. We observed a significant reduced expression of GGT1 in the ileum and caecum of patients with PSC/UC relative to both healthy individuals and patients with UC. This finding provides a compelling hypothesis linking liver function to intestinal redox homeostasis and colon cancer development. Indeed, elevated serum GGT is often used as a marker of cholesterol and liver dysfunction. Its role in liver function is also evident from recent observations of single nucleotide variation at the GGT1 locus being associated with PSC.

Our observations extend altered GGT function to the intestine. GGT1 is an important enzyme in the glutathione salvage pathway, degrading extracellular glutathione into its amino acid constituents for cellular uptake. Once inside the cell, glutathione is reformed and participates in the reduction of reactive oxygen species. Using publicly available data, we have shown that GGT1 positive cells are present in the epithelial compartment of the small and large intestine. The predominance of GGT1 positive cells in enterocytes appears to lead to an increase in GGT1 positive cells in the ileum relative to the colon and rectum. This is noteworthy, as it suggests developmental differences between tissue sites that we believe may relate to the availability of liver-derived factors. We speculate that reduced bile flow has an impact on GGT1 expression in the intestine. In this regard it is of interest that increased serum bilirubin has been associated with an increased risk of cholangiocarcinoma in patients with PSC. Whether this is also true for colon cancer development remains to be explored. Together these data suggest a new hypothesis of colon cancer development in PSC/UC that involves a liver dysfunction-induced deficit in the ability of intestinal epithelial cells to combat oxidative stress in the context of inflammation. We speculate that this may be a key driver of colon cancer development through increased DNA damage or metabolic dysfunction as has been shown for ChRCC.

In addition to host factors in PSC/UC, the microbiome has become increasingly recognised as an important contributory
factor. Indeed, members of the microbiome are important mediators of the gut-liver axis through their ability to metabolise bile acids that reach the intestine from the liver. Multiple studies have provided links between the bacterial component of the microbiome from both stool\textsuperscript{12,13,79} and mucosa\textsuperscript{45,77} with PSC/UC. Furthermore, mouse studies have provided a potential causal link between PSC-associated microbial consortia and Th17-driven immunopathology in both the gut and the liver\textsuperscript{15}. We did not observe any striking effects of the bacterial component of the microbiome across three tissue locations in patients with PSC/UC. After accounting for the presence of contaminants in our data set, we observed a single genus, \textit{Parabacteroides}, as being significantly reduced in the rectum of patients with PSC/UC. This observation is consistent with a similarly reduced relative abundance of \textit{Parabacteroides} in neonatal cholestatic disease\textsuperscript{77}. While not consistently observed in PSC, there has also been a report of increased \textit{Parabacteroides} in the stool of patients with PSC\textsuperscript{78}. While contradictory in direction to the current study, differences in sampling method may be a relevant factor. Further work is required to fully disentangle genuine microbiome signals from these studies in the context of multiple influencing factors that may vary across study cohorts.

Finally, we show limited evidence for associations between genera abundances and expression of host genes. Previous studies have observed relationships between bacterial abundance and host gene expression\textsuperscript{79} and we cannot discount that we simply did not have the power to detect robust associations in our relatively small cohort, especially given the sparseness of the data under consideration. An important observation from our data is the potential of CLR-transformed zero counts to drive host-microbiome associations. Zero counts are a known issue when using log-ratio transforms for microbiome data\textsuperscript{45,79} and while the ALDEx2 CLR transform deals well with zero counts we could not exclude that spurious correlations were introduced due to zero-count transformation. Using non-zero data for correlation analysis will likely have impacted our ability to identify robust host-microbiome associations and is a limitation of this study.

In conclusion, we have shown tissue-dependent gene expression differences across intestinal diseases and identified an association between reduced \textit{GGT1} expression in the ileum and caecum of patients with PSC/UC that we hypothesise to link liver function with colorectal cancer risk in the context of inflammatory bowel disease.

**Data availability**

**Underlying data**


16S rRNA sequencing data have been deposited in the

EMBL-EBI: 16S rRNA sequencing data. Accession number PRJEB40737; [https://identifiers.org/ena.emb|PRJEB40737](https://identifiers.org/ena.emb|PRJEB40737).

Extended data

Analysis code available from: [https://github.com/nickilott/PSC_UC_Host_Microbiome](https://github.com/nickilott/PSC_UC_Host_Microbiome)

Archived analysis code at time of publication: [https://doi.org/10.5281/zenodo.4808793\textsuperscript{46}](https://doi.org/10.5281/zenodo.4808793).

License: MIT


This project contains the following extended data:

- Figure S1 (Association of demographic, medication use, inflammation parameters and liver function with disease status.)
- Figure S2 (Relative abundance of a known caecal microbiome community across serially diluted caecal samples.)
- Figure S3 (Schematic overview of generating host-microbiome correlation networks.)
- Figure S4 (Correlation network generated before accounting for a role of zero counts in genus CLR-transformed counts. Zero counts are shown to impact observed associations between host gene expression and genus CLR.)
- Table S1 (DESeq2 differential expression results in an LRT analysis between tissue sites in healthy individuals.)
- Table S2 (Cluster assignments of genes to tissue-defining modules using dynamicTreeCut.)
- Table S3 (GO:BP enrichment results for genes in cluster 1.)
- Table S4 (GO:BP enrichment results for genes in cluster 2.)
- Table S5 (GO:BP enrichment results for genes in cluster 3.)
- Table S6 (GO:BP enrichment results for genes in cluster 4.)
- Table S7 (Transcription factor enrichment results for genes in cluster 1.)
- Table S8 (Transcription factor enrichment results for genes in cluster 2.)
- Table S9 (Transcription factor enrichment results for genes in cluster 3.)
- Table S10 (Transcription factor enrichment results for genes in cluster 4.)
- Table S11 (DESeq2 differential expression results for Tissue: Ileum, contrast: PSC/UC vs. healthy.)
- Table S12 (DESeq2 differential expression results for Tissue: Ileum, contrast: UC vs. healthy.)
- Table S13 (DESeq2 differential expression results for Tissue: Caecum, contrast: PSC/UC vs. healthy.)
- Table S14 (DESeq2 differential expression results for Tissue: Caecum, contrast: UC vs. healthy.)
- Table S15 (DESeq2 differential expression results for Tissue: Rectum, contrast: PSC/UC vs. Healthy)
- Table S16 (DESeq2 differential expression results for Tissue: Rectum, contrast: UC vs. healthy.)
- Table S17 (DESeq2 differential expression results for Tissue: Ileum, contrast: PSC/UC vs. UC)
- Table S18 (DESeq2 differential expression results for Tissue: Ileum, contrast: UC vs. PSC/UC)
- Table S19 (DESeq2 differential expression results for Tissue: Rectum, contrast: PSC/UC vs. UC)
- Table S20 (Putative contaminant ASVs.)
- Table S21 (DESeq2 genus differential abundance results for Tissue: Ileum, contrast: PSC/UC vs. healthy.)
- Table S22 (DESeq2 genus differential abundance results for Tissue: Ileum, contrast: UC vs. healthy.)
- Table S23 (DESeq2 genus differential abundance results for Tissue: Caecum, contrast: PSC/UC vs. healthy.)
- Table S24 (DESeq2 genus differential abundance results for Tissue: Caecum, contrast: UC vs. healthy.)
- Table S25 (DESeq2 differential expression results for Tissue: Rectum, contrast: UC vs. healthy.)
- Table S26 (DESeq2 differential expression results for Tissue: Rectum, contrast: UC vs. healthy.)
- counts.zip (Processed counts tables for RNA-seq and 16S amplicon sequencing data.)
- annotations.zip (Annotation files for reproducing data analysis.)
- metadata.zip (Metadata files for reproducing data analysis.)

Reporting guidelines

Data are available under the terms of the Creative Commons Attribution 4.0 International license (CC-BY 4.0).

Acknowledgements
We thank the Oxford Genomics Centre at the Wellcome Centre for Human Genetics (funded by Wellcome Trust grant reference 203141/Z/16/Z) for the generation and initial processing of sequencing data. The views expressed are those of the authors and not necessarily those of the National Health Service, the National Institute for Health Research, or the Department of Health. We acknowledge the contribution of the BRC Gastrointestinal Biobank (11/YH/0020 and 16/YH/0247), which is supported by the National Institute for Health Research, Oxford Biomedical Research Centre (BRC). The Translational Gastroenterology Unit (TGU) Investigators consist of multiple researchers at the University of Oxford. These researchers are Dr Philip Allan, Dr Timothy Ambrose, Dr Adam Bailey, Professor Ellie Barnes, Dr Elizabeth Bird-Liebermann, Dr Jan Bornschein, Dr Barbara Braden, Dr Oliver Brain, Dr Jeremy Cobbold, Dr. Emma Culver, Professor Chris Cunningham, Professor James East, Dr Lucy Howarth, Professor Paul Kloenerman, Professor Simon Leedham, Dr Rebecca Palmer, Dr Michael Pavlides, Professor Fiona Powrie, Dr Astor Rodrigues, Professor Jack Satsangi, Professor Simon Travis and Dr Alissa Walsh. We would like to thank them, as well as Dr Kate Lynch, for giving up their time and resources to build research within the TGU. The authors thank all volunteers and patients who took part in this study.

References


PubMed Abstract | Publisher Full Text


PubMed Abstract | Publisher Full Text


PubMed Abstract | Publisher Full Text | Free Full Text


PubMed Abstract | Publisher Full Text


PubMed Abstract | Publisher Full Text | Free Full Text


PubMed Abstract | Publisher Full Text


PubMed Abstract | Publisher Full Text


PubMed Abstract | Publisher Full Text | Free Full Text


PubMed Abstract | Publisher Full Text


PubMed Abstract | Publisher Full Text | Free Full Text


PubMed Abstract | Publisher Full Text


PubMed Abstract | Publisher Full Text | Free Full Text


PubMed Abstract | Publisher Full Text


PubMed Abstract | Publisher Full Text | Free Full Text


PubMed Abstract | Publisher Full Text


PubMed Abstract | Publisher Full Text | Free Full Text


PubMed Abstract | Publisher Full Text | Free Full Text


PubMed Abstract | Publisher Full Text


PubMed Abstract | Publisher Full Text | Free Full Text


PubMed Abstract | Publisher Full Text | Free Full Text


PubMed Abstract | Publisher Full Text | Free Full Text


PubMed Abstract | Publisher Full Text | Free Full Text


PubMed Abstract | Publisher Full Text | Free Full Text


PubMed Abstract | Publisher Full Text | Free Full Text


PubMed Abstract | Publisher Full Text | Free Full Text


PubMed Abstract | Publisher Full Text | Free Full Text


PubMed Abstract | Publisher Full Text | Free Full Text


PubMed Abstract | Publisher Full Text | Free Full Text


PubMed Abstract | Publisher Full Text | Free Full Text