Manuscript Doi: 10.1093/ecco-jcc/jjaa236

Ileal transcriptomic analysis in paediatric Crohn's disease reveals *IL17*- and *NOD*-signalling expression signatures in treatment-naïve patients and identifies epithelial cells driving differentially expressed genes

James J Ashton[1,2], Konstantinos Boukas[3], James Davies[4], Imogen S Stafford[1,5], Andres F. Vallejo[4], Rachel Haggarty[6], Tracy AF Coelho[2], Akshay Batra[2], Nadeem A Afzal[2], Bhumita Vadgama[7], Anthony P. Williams[3,5], R Mark Beattie[2], Marta E Polak[4,8], and Sarah Ennis[1]

1. Department of Human Genetics and Genomic Medicine, University of Southampton, Southampton, UK
2. Department of Paediatric Gastroenterology, Southampton Children's Hospital, Southampton, UK
3. Wessex Investigational Sciences Hub, University of Southampton Faculty of Medicine, Southampton General Hospital, Southampton, UK
4. Clinical and Experimental Sciences, Sir Henry Wellcome Laboratories, Faculty of Medicine, University of Southampton, Southampton, UK
5. Institute for Life Sciences, University of Southampton, Southampton, UK
6. NIHR Southampton Biomedical Research Centre, University Hospital Southampton, Southampton, UK
7. Department of Paediatric Histopathology, Southampton Children's Hospital, Southampton, UK
8. Institute for Life Sciences, University of Southampton, Southampton, UK

Correspondence to-

Professor Sarah Ennis

Human Genetics & Genomic Medicine

University of Southampton

Duthie Building (Mailpoint 808)

Southampton General Hospital

Southampton

SO16 6YD

Tel: +44 (0)23 8079 8614

s.ennis@southampton.ac.uk

Conflicts of interest

The authors declare no conflicts of interest

Author contributions

JJA, RMB and SE conceived the study. Patients were recruited by JJA and RH. Patient samples were acquired by JJA, RH, TAC, AB, NAA and RMB. Samples were processed and sequenced by JJA, KB and JD. Analyses were performed by JJA, KB, JD and ISS, under the guidance of AFV, MP, AW and SE. BV performed histological analysis for all patients. JJA wrote the manuscript with help from all authors. All authors approved the final manuscript prior to submission.

Data Availability Statement

All data generated in this project are available via the GEO repository- GSE153974 and GSE153866

Abstract

Background/Aims- Crohn's disease (CD) arises through host-environment interaction. Abnormal gene expression results from disturbed pathway activation or response to bacteria. We aimed to determine activated pathways and driving cell types in paediatric CD.

Methods- We employed contemporary targeted autoimmune RNA sequencing, in parallel to single-cell sequencing, to ileal tissue derived from paediatric CD and controls. Weighted-gene-co-expression-network-analysis (WGCNA) was performed and differentially expressed genes (DEGs) were determined. We integrated clinical data to determine co-expression modules associated with outcomes.

Results- Twenty-seven treatment-naive CD (TN-CD), 26 established-CD patients and 17 controls were included. WGCNA revealed a 31-gene signature characterising TN-CD patients, but not established-CD, or controls. The *CSF3R* gene is a hub within this module and is key in neutrophil expansion and differentiation.

Antimicrobial genes including *S100A12* and the calprotectin subunit *S100A9* were significantly upregulated in TN CD compared to controls ($p=2.61x10^{-15}$ and $p=9.13x10^{-14}$, respectively) and established-CD (both p=0.0055). Gene-enrichment analysis confirmed upregulation of the IL17-, NOD- and Oncostatin-M-signalling pathways in TN-CD patients, identified in both WGCNA and DEG

analyses. An upregulated gene-signature was enriched for transcripts promoting Th17-cell differentiation and correlated with prolonged time to relapse (correlation-coefficient-0.36, p=0.07).

Single-cell sequencing of TN-CD patients identified specialised epithelial cells driving differential expression of *S100A9*. Cell groups, determined by single-cell gene-expression, demonstrated enrichment of IL17-signalling in monocytes and epithelial cells.

Conclusion- Ileal tissue from treatment naïve paediatric patients is significantly upregulated for genes driving IL17-, NOD- and Oncostatin-M-signalling. This signal is driven by a distinct subset of epithelial cells expressing antimicrobial gene transcripts.

Introduction

Paediatric-onset Crohn's disease is a heterogenous condition characterised by chronic, relapsing and remitting inflammation, largely of the intestinal tract. Paediatric-onset Crohn's disease has a greater genetic contribution to pathogenesis compared to adult-onset disease, with multiple genes and pathways implicated in the inflammation observed in the condition[1]. These genes are largely centred on innate and adaptive immune pathways, cytokine signalling pathways, and bacterial recognition and response pathways[2]. Recently Mendelian causes of inflammatory bowel disease (IBD) have given additional insights into causes and risk factors for polygenic disease, with variation in a number of genes including *IL10* pathway, *NOD2* and NADPH oxidase complex genes being implicated in both forms of disease[3–5]. Non-Mendelian forms of Crohn's disease may present with similar phenotypic appearance. However, it is becoming increasingly clear that individual patients are likely to have a specific molecular diagnosis related to their underlying genetic variation. This may present through either a limited number of genes (oligogenic IBD) or through interaction of many genes (polygenic IBD), often resulting in perturbation of inflammatory pathways common to all genetic causes[2]. The ability to determine this genetic signature within an individual patient will bring new opportunities for predicting disease outcome and personalisation of therapy[6].

RNA sequencing allows identification of abnormal gene expression, and specific gene signatures, which are associated with disease subtypes. This information provides insight into the biological processes underlying pathways driving inflammation. Previous studies have identified differentially expressed genes, including *OSM* and *TREM1,* associated with disease onset, treatment response and have been able to predict disease course [7–9]. Whilst there is clear utility in determining markers of disease in blood, insights from non-gastrointestinal tissues are likely sub-optimal to elucidate drivers of intestinal inflammation[10]. In addition, bulk RNA sequencing, where all cell types resident within a single biopsy sample are assessed concurrently, may to fail to sequence lowly expressed genes, with

reads being overwhelmed by housekeeping transcripts that provide little biological insight[11]. In contrast, contemporary efforts at targeted and single-cell sequencing in cancer have revealed novel pathways and genes associated with disease, and provided clinical diagnostic value[12–14]. Furthermore, integration of targeted RNA sequencing, single cell sequencing and clinical outcome data provides the opportunity for improved molecular profiling of patients, garnering understanding of the specific cells driving inflammatory pathways whilst simultaneously using RNA gene signatures to stratify patients. Recently, single cell analysis in 22 adult Crohn's disease patients determined a specific transcriptomic module from cells derived from the lamina propria, which was reproducibly found in bulk RNA sequencing and was associated with failure to respond to anti-TNF therapy[15].

In this study we apply cutting-edge autoimmune targeted RNA sequencing of ileal biopsy tissue from a cohort of paediatric Crohn's disease patients. We utilise these data to characterise patients by underlying gene transcription signatures, identify differentially expressed genes impacting on signalling pathways in treatment naïve and established disease patients. We integrate single cell RNA sequencing performed on a subset of patients to determine cell populations driving specific gene expression.

Methods

Paediatric IBD patients were recruited through from the Paediatric Gastroenterology service at the Southampton Children's Hospital. All patients were diagnosed under the age of 18 years, according to the modified Porto criteria[16]. Two patient populations were recruited, the first consisted of patients referred to paediatric gastroenteology with a suspected diagnosis of IBD, recruited prior to diagnostic endoscopy. Patients who were diagnosed with Crohn's disease following successful ileoscopy, were labelled as the treatment-naïve (TN) Crohn's disease group. Children who had a normal endoscopy and were not diagnosed with IBD or any other gastrointestinal pathology, were included in a control group. The control group was followed-up for a minimum of 6 months to

confirm there was no subsequent diagnosis of IBD. The second population consisted of patients with established (ED) Crohn's disease undergoing routine endoscopy for reassessment, termed the established disease group (figure 1 and supplementary methods).

Ethical approval

The study has category A ERGO II ethics approval (30630) and a REC approval from Southampton and South West Hampshire Research Ethics Committee (09/H0504/125). All patients and families provided informed consent at recruitment.

RNA sequencing of terminal ileal biopsies

Sample acquisition, processing and storage

Terminal ileal biopsies were obtained during endoscopy and immediately placed into a cryovial containing 1ml of RNAlater (Sigma Aldrich), and frozen at -80°C within 30 minutes from collection. The diameter of each biopsy was an average of 2.5mm (range 1-4.5mm) with the mean volume of 27mm^3 (equivalent to 30mg).

RNA extraction

Biopsies were homogenized using Qiagen TissueLyser™ and RNA was extracted using Promega Maxwell RSC™ (simplyRNA tissue) and frozen at -80°C. RNA quality was assessed using the Agilent Bioanalyzer™ (RNA Nano). RIN values and RNA concentrations are available as supplementary data 1.

Targeted RNA sequencing

The contemporary HTG EdgeSeq Autoimmune Panel was used to measure mRNA expression levels in 2002 autoimmune genes, including inflammatory bowel disease[17]. Briefly, RNA samples were thawed, diluted and 25ng of RNA per sample was plated in 96 wells and loaded onto the HTG

EdgeSeq instrument. HTG fully automated nuclease target protection chemistry includes hybridization of mRNA to target-specific Nuclease Protection Probes (NPPs), addition of S1 nuclease to digest excess NPPs and non-hybridized RNA followed by heat denaturation of S1. The processed samples contain 5' and 3'-end wing sequences and were used as a template for PCR reactions with primers complementary to the "wings" which also contain barcode sequences and common adaptors required for cluster generation and sequencing (P5, P7). Libraries were cleaned-up following a standard clean-up procedure (AMPureXP, PEG8000) and quantified with qPCR using KAPA Library quantification (Roche) kit. Libraries were normalized accordingly, were pooled in a 3pM concentration and were loaded on the Illumina NextSeq500 for single read deep sequencing (Read1: 50bp, Index1: 6bp, Index2: 6bp).

RNA data processing

Sequencing output basecalls were converted into FASTQ and demultiplexed using module bcl2fastq2/2.18 on IRIDIS HPS (University of Southampton), adding the option --barcode-mismatches 0. FASTQ files were parsed on HTG EdgeSeq parser (v 5.2.823) constructing a gene expression count matrix for each gene and each patient. They were merged to form a single output file containing all genes and all counts. Downstream analyses of RNA data were performed using HTG Reveal, a web-based, GDPR-compliant data analysis suite. Gene counts were normalized using quantile normalization (QN) based on best practice guidelines previously applied with HTG EdgeSeq targeted immuno-oncology panel[18,19].

RNA sequencing data quality assessment and analysis

Quality control of RNA sequencing data was performed in line with recommendations from HTG, to satisfy cut-offs for sample quality, sufficient read depth and minimal expression variability across probes. Differential expression was assessed using DESeq2 package (Python, within Reveal software, 2020 version)[20]. Gene-co-expression networks enable regulatory hubs and gene-gene associations to

be determined. CEMiTool (2020 version) was used to assess weighted gene co-expression networks within normalized data, and to determine modules and hub regulatory genes observed in different categorical groups[21]. Gene-Gene interactions within co-expressed genes were determined using the HitPredict database[22]. WGCNA (R package in R studio version 1.2.1335) was used to establish gene co-expression modules and assess correlation between continuous clinical outcome variables[23].

We assessed for enrichment of genes in specific WGCNA modules, and differentially expressed genes (DEGs), in specific pathways using ToppFun[24], EnRichR[25] and BioPlanet[26]. Statistical analysis was performed using Reveal software and SPSS (v25, IBM).

Raw read count matrix and metadata spreadsheet are available in the NCBI Gene Expression Omnibus (GSE153974).

Single-cell transcriptomic analysis

Fresh ileal tissue biopsies from two Crohn's disease patients were digested within 10 minutes of biopsy procedure, cells were disaggregated, slow frozen (10%DMSO) and thawed just before use.

Co-encapsulation of single cells with genetically encoded beads was performed following the Drop-seq pipeline[27,28]. Optimised microfluidics parameters were used, ensuring the generation of single-cell/single barcoded Bead SeqB (Chemgenes, USA) encapsulation events. Following encapsulation, ~4500 STAMPS (beads exposed to a single cell) from 1.2 ml of cell suspension were generated. 1000 STAMPS for each biopsy were taken further for library preparation (High Sensitivity DNA Assay, Agilent Bioanalyser, 12 peaks with the average fragment size 500 bp). Prepared libraries were run on an Illumina NextSeq (1×10^5 reads/cell).

Alignment, read filtering, barcode and Unique Molecular Identifier (UMI) counting were performed using STAR[29]. For gene filtering, genes detected in less than 10 cells were excluded. Subsequent data analyses were run using the python-based Scanpy[30] with R (3.6.0). Background empty barcodes were

identified and removed using EmptyDrops[31]. Cells of low quality with high fraction of counts from mitochondrial genes (20% or more), which indicates stressed or dying cells, were removed. Data was normalised using Scran[32]. A single-cell neighbourhood graph, with data integrated from separate tissue samples, was computed using BBKNN[33]. Data were visualised using Uniform Manifold Approximation and Projection (UMAP), with the Leiden algorithm used to identify cell clusters[34].

Cell type annotation was performed using SingleR (database: BlueprintEncodeData) [35]. Markers genes for cell clusters were identified using a T-test within Scanpy[30].

Raw sequencing data are available in the NCBI Gene Expression Omnibus (GSE153866).

Clinical data integration

Clinical outcome data were collected on all treatment naïve patients. Time to first clinical relapse was used as the primary clinical outcome measure. We defined relapse as requirement for repeated steroid course, or exclusive enteral nutrition, or step-up in immunomodulation or biologic therapy, after completion of initial induction. All patients had entered remission following induction therapy. Patients that had not relapsed at most recent follow-up were arbitrarily assigned their time from diagnosis to most recent follow-up for analysis.

Histological evidence of ileitis was recorded for all patients. Clinical biopsies were taken concurrently and examined by a paediatric histopathologist to determine the presence of inflammation within the ileum at the time of endoscopy.

Results

Ninety-one patients with ileal biopsies were recruited to the study. Confirmation of diagnosis by Porto criteria resulted in 70 patients being included: 27 TN Crohn's disease patients; 17 controls; and 26 ED Crohn's disease. Twenty-one patients that were diagnosed with IBDU or ulcerative colitis

following endoscopy were excluded. A single ED sample failed targeted RNA quality control (parameter two, sample had low variation of counts across probes) leading to exclusion. Patient characteristics can be seen in table 1. Ileal biopsies taken from the TN Crohn's disease patients were from inflamed areas in 74% of cases, for established disease patients only 27% had biopsies were from inflamed areas.

Targeted RNA sequencing of 2002 autoimmune genes

A thirty-one gene module characterises treatment-naïve Crohn's disease patients

Gene modules associated with TN patients, controls and ED patients were established. Three gene modules were identified containing 104, 47 and 31 genes respectively (supplementary data 2). Module 1 co-expression was significantly increased in controls (NES 1.95, p=0.0004) and significantly decreased in TN (NES -1.7, p=0.0015) and ED (NES -2.35, p=0.0012) patients. Whilst module 2 contained 47 co-expressed genes it was not significantly associated with any patient group. Module 2 was significantly enriched for cell metabolic processes, including PPAR signalling ($p=1.22x10^{-7}$) and protein digestion and absorption (p= 0.006055). Module 3, containing 31 genes, was significantly upregulated in TN patients (normalised expression score, NES 3.07, p=0.0006) and downregulated in controls (NES -2.73, p=0.0004), figure 2A-B and supplementary table 1. Module 3 did not correlate with ED patients.

The treatment-naïve gene co-expression module is associated with upregulation of Oncostatin-M and NOD-signalling pathways

Thirty-three pathways were significantly associated with module 3 genes following multiple testing correction (supplementary data 3). The most implicated pathway was the Oncostatin-M (OSM) signalling pathway ($adj\text{-}p=4.47x10^{-22}$), upregulation of which was seen in treatment-naïve patients. OSM signalling results in activation of proinflammatory pathways including *JAK/STAT3*, *MAPK*, and

PI3K. Interestingly, given the key role of NOD2 in Crohn's disease pathogenesis, activation of the NOD-signalling pathway was also significantly enriched for in module 3 (adj-p=0.0008).

CSF3R appears to act as a regulatory hub within the treatment-naïve module

In order to assess regulatory hub genes within the module 3 network we performed an interaction network analysis. This revealed six genes with >3 hub interactions within the module, figure 2C. Of these six genes *CSF3R* was also co-expressed within the network with *AQP9*, figure 2C. *CSF3R* is the receptor for colony stimulating factor 3. The related pathway functions to control expansion, differentiation and role of neutrophils, with highly deleterious variants in *CSF3R* resulting in congenital neutropenia.

S100A9 (Calprotectin subunit) and S100A12 antimicrobial genes are significantly upregulated in treatment-naïve patients

Differential gene expression was assessed between TN patients, controls and ED patients using DESeq2 (supplementary data 4 and 5). Following multiple testing correction, 342 genes were significantly differentially expressed between TN patients and controls, 259 of which were upregulated in TN patients (logFC >1.15, FDR <0.05, figure 3A). The five most significant upregulated DEGs in TN patients were *S100A12* (fold change 32.5, adj-p= 2.6×10^{-15}), *CXCL8* (IL8)(fold change 20.2, adj-p= 5.5×10^{-15}), *S100A9* (fold change 12.6, adj-p= 9.1×10^{-14}), *FCGR3A/B* (fold change 7.3, adj-p= 5.0×10^{-13}), and *IL1RN* (fold change 8.1, adj-p= 3.9×10^{-12}). The difference between TN and ED patients was less marked, whereby just 12 genes were significantly upregulated in TN patients compared to ED patients (Figure 2D, logFC >1.15, FDR <0.05). The five most significantly upregulated genes were *CSF3R* (fold change 2.5, adj-p= 0.0055), *IL1RN* (fold change 3.0, adj-p= 0.0055), *S100A9* (fold change 3.6, adj-p= 0.0055), *S100A12* (fold change 4.8, adj-p= 0.0055) and *AQP9* (fold change 3.6, adj-p= 0.03).

Treatment naïve Crohn's disease is characterised by elevated IL17- and NOD-signalling.

We utilised gene enrichment analysis to assess for pathways associated with TN patients using genes implicated by *both* WGCNA and differential gene expression analysis. The IL17- and NOD-signalling pathways were recurrently implicated across multiple gene enrichment databases (Supplementary table 2). These pathways have previously been implicated in Crohn's disease pathogenesis, and more so, are a biologically plausible explanation for a chronic inflammatory process within the intestine[36,37].

Gene expression differences are not driven solely by inflamed tissue

To ensure differences between groups were not driven solely by active inflammation we conducted analysis on inflamed vs. non-inflamed tissue. Whilst there were differentially expressed probes between inflamed and non-inflamed biopsies (total DEGs=185), the most differentially expressed genes were different to those observed when comparing TN CD and controls (Supplementary figure 1).

Gene expression in NOD-signalling pathway clusters Crohn's disease patients distinctly from controls

NOD2 is the most heavily implicated gene in Crohn's disease pathogenesis, with variation in multiple interacting genes, including *XIAP, RIPK2* and *ATG16L1,* described as increasing risk of Crohn's disease[2]. Transcripts from genes within the NOD-signalling pathway were amongst the most significantly enriched in treatment naïve patients. We hypothesised that aberrant NOD-signalling gene expression could be used to classify patients from controls. Utilising a list of 95 genes in the NOD-signalling pathway, curated by the HTG platform, we performed hierarchical clustering of all patients (quantile normalised data, average distance clustering), figure 4. Three broad clusters were formed, with 8 patients remaining outliers. All but three of the controls grouped together in cluster 3, characterised by low *CXCL8 (IL8), CXCL2 and CASP5* expression. In contrast clusters 1 and 2 had

increased expression of pro-inflammatory *CXCL1* and *STAT1. NOD2* expression itself did not appear to differ between groups.

Clinical data integration

Th17-cell differentiation gene module associated with prolonged time to relapse

Using WGCNA, we assessed whether co-expressed gene modules were predictive for patient prognosis. The number of days from diagnosis to relapse were entered as a continuous variable. Six of the 27 treatment naïve Crohn's disease patients had not relapsed at the time of analysis. We took the conservative approach of setting their time to relapse as the number of days from diagnoses to most recent follow up. A co-expression module ('blue', supplementary data 6) characterised by 55 significantly upregulated genes (p= <0.05), was positively correlated with time to relapse (correlation coefficient 0.36, p=0.07) (Supplementary figure 2). The 55 genes comprising this module were significantly enriched for Th17 cell differentiation (KEGG, adjusted p value 9.21×10^{-11}), indicating patients with longer duration of remission had increased activation of Th17 cell differentiation pathways, supplementary data 7.

In order to determine if any DEGs were associated with early relapse we stratified the 27 treatment naïve CD patients into those that relapsed within 16 weeks (n =14) and those that either relapsed after 16 weeks or did not relapse (n =13). Sixteen weeks was chosen to split the cohort into two equal groups to increase the power to identify DEGs. Expression of the *PI3* gene, an antimicrobial peptide expressed in response to lipopolysaccharide and IL17-signalling, was significantly upregulated in patients with early relapse (fold change = 2.13, adjusted p=0.0358, supplementary table 3)[38].

Single-cell sequencing of treatment naïve Crohn's disease patients

Specialised gastrointestinal epithelial cells drive differential expression of S100A9, a key molecule in IL17 signalling pathway.

Single cell RNA sequencing of ileal biopsies from two treatment naïve Crohn's disease patients (patient IDs SOPR0472 and SOPR0476) were undertaken in order to identify distinct cell populations driving the enrichment of specific biological pathways. After filtering, a total of 1458 cells (SOPR0472 =710 cells, SOPR0476=748 cells) and 4,731 genes were used for analyses. ScanPy UMAP dimensionality reduction of the 1458 cells integrated from both patient biopsies (SOPR0472=710, SOPR0476=748 cells) was performed (Figure 5A), followed by annotation of cell populations as CD8+ effector memory T cells (CD8+ Tem), memory B-cells, monocytes, epithelial cells and plasma cells, contributed by cells from each sample (Figure 5A, 4B). We identified 9 distinct clusters within the SingleR annotated populations (Figure 5C), with each cluster defined by the expression of unique marker genes (Figure 5D), supplementary data 8. Cells were assigned to clusters based on their overall transcription profiles and confirmed by further annotation by Enrichr CellType (supplementary data 9). Interestingly, the 2 genes that most strongly defined cluster 7 as epithelial cells included both calprotectin subunits, *S100A8* and *S100A9*. Interestingly, the *S100A9* gene was one of the top DEGs between TN patients and controls. *S100A8* is the other calprotectin subunit but as this transcript was not probed in the targeted panel we could not compare its expression between methodologies. Gene ontology analysis of the 50 genes that define cluster 7 (n=50) revealed associations with inflammatory immune response processes, including an epithelial defence response to bacterium (adj-p=$2.2x10^{-5}$).

Specific cell populations drive gene expression seen in treatment naïve patients

The nine discrete cell populations were interrogated for their relative expression of genes identified in the gene module 3 that characterised TN CD patients (Figure 2A). The cell populations showing elevated expression of module 3 genes included cells identified as monocytes (cluster 5- *CCL4, CXCL3, IL1RN, IL8* and *PLAUR*) and epithelial cells in cluster 7 (*S100A9*), cluster 4 (*LCN2, MMP1* and *PLAUR*) and cluster 2 (*IL8, LCN2, MMP1, PI3* and *PLAUR*)(Figure 5E).

Enrichment of Il17-signalling was observed in monocytes and specialised epithelial cells

We hypothesised that specific cell populations were driving the pathways implicated by targeted bulk RNA sequencing. IL17-signalling genes were enriched within the specialised epithelial cells (cluster 7), and in monocytes (cluster 5). Significant enrichment for the IL17 pathway (adj-p=0.016) in cluster 7 was largely attributed to elevated expression of *S100A7, S100A8* and *S100A9*. Cluster 5 monocytes markers (n=50) were enriched for genes involved in IL17 signalling pathway (adj-p=0.0012); due to elevated expression of *CXCL8* (IL8), *CXCL3, IL1B, NFKBIA, HSP90B1,* as well as the broad pathways of 'response to cytokines' (adj-p=$1.1x10^{-9}$) and the 'inflammatory response' (adj-p=$2.6x10^{-6}$).

Discussion

We present data utilising a targeted autoimmune panel for the first time in IBD. These data demonstrate an ileal gene expression signature identified through both WGCNA and differential gene expression analysis, characterised by NOD- and IL17-signalling, and specific to treatment-naïve Crohn's disease patients. Parallel single cell analysis identified specialised epithelial cells driving differential expression of several genes, including the calprotectin subunits (*S100A8*/*S100A9*). Enrichment of IL17-signalling genes was observed in both this epithelial cell cluster as well as in a distinct subset of monocytes. Finally, we identify a gene module characterised by the Th17 cell

differentiation pathway that is observed in patients with increased time to relapse following diagnosis.

Our data confirm and corroborates findings from a number of previous studies, implicating *OSM*, *CXCL8* and *AQP9* as upregulated DEGs in IBD patients compared to controls[7,8]. We also provide tissue-specific evidence for several upregulated genes that have previously been observed in blood, including *TREM1*[9]. We confirm the findings of Haberman *et al*, who previously detailing paediatric Crohn's disease ileal transcriptomic signatures. We identify a large number of overlapping genes from their 'core ileal Crohn's disease' expression module, including upregulation of *CXCL5*, *IL8* and *S100A9*[8]. These genes, including IL8, OSM and TREM1 are either pro-inflammatory cytokines, or receptors triggering downstream inflammatory signalling and have been linked to inflammatory processed in autoimmune disease[7,9,39]. In addition, we identify the hub gene, CSF3R, a regulator of neutrophil differentiation, which appears to be key in controlling expression of a number of genes in treatment-naïve Crohn's disease. This gene has been previously associated with congenital neutropenia but this is the first time it has been specifically implicated in IBD pathogenesis. This is despite the fact that a number of primary immunodeficiencies are recognised to present with IBD-like phenotypes[40].

Stratification of patients based on differential gene expression provides the opportunity to predict treatment response and patient prognosis. Recently the RISK cohort from North America has been used to develop prediction algorithms, specifically utilising integration of ileal transcriptome profiles into a multifactorial risk score for stricturing disease, identifying an extracellular matrix gene signature associated with stricturing disease[41]. Haberman *et al* also created a model, including *APOA1* gene expression, able to predict 6-month steroid free remission[8]. Here we identify a gene module, characterised by Th17 cell differentiation, associated with prolonged time to relapse. Once

the results are replicated in external cohorts, transitioning these models into clinical practice will be key for personalising therapy in patients.

Th17 cells and IL17 signalling are of great interest in IBD pathogenesis. Differentiated Th17 cells produce several effector IL17 cytokines, promoting inflammation and mucosal pathogen clearance[37]. We implicate ileal activation of IL17 signalling within TN Crohn's disease patients, compared to both controls and ED patients. Several previous studies have not identified increased serum IL17 in Crohn's disease patients, however increased IL17 levels have been detected in affected tissues[42,43]. We replicate these findings in paediatric patients and for the first time identify a specialised epithelial cell cluster and a monocyte cell cluster appearing to drive the IL17-signalling. We hypothesise that the epithelium appears to be a target tissue for IL17-signalling, resulting in heightened proinflammatory and anti-microbial response within this specific population of cells, characterised by secretion of calprotectin (*S100A8/S100A9*). This pro-inflammatory effect of IL17-signalling on epithelial cells has been observed in colonic epithelial cell cultures, with upregulation of CXCL8 and CXCL1 promoting neutrophil chemotaxis[44]. Our data identify distinct cell clusters driving these processes, whilst targeting RNA sequencing replicates similar expression profiles within the ileum of treatment naïve patients.

Previously, Martin *et al* identified an expression module in ileal Crohn's disease, characterising IgG plasma cells, mononuclear phagocytes, activated T cells and stromal cells[15]. Through single cell sequencing of ileal tissue derived from two TN patients we were able to identify a novel group of epithelial cells driving differential expression of the calprotectin complex (*S100A8/S100A9*), which was echoed in the targeted sequencing of all patients. Typically, it has been thought that most calprotectin is derived from colonic neutrophils, with small bowel inflammation less reflected in faecal sampling[45]. Our data indicate that specialised ileal epithelial cells highly express *S100A8/S100A9* in Crohn's disease patients, driving the differential gene expression between

patients and controls. It appears that these ileal epithelial cells show an increased expression of these proteins under inflammatory conditions. Interestingly, previous transcriptomic analysis of only intestinal epithelial cells of paediatric Crohn's disease patients did not identify any differentially expressed genes correlated to active inflammation[46]. In their article, Howell *et al* describe several genes, including *DEFA5*, *DEFA6, LYZ*, *PLA2G2A*, *CD40*, and *CD44*, that were differentially expressed between controls and TN Crohn's disease patients, concluding that there was minimal molecular impact of disease on the epithelial cells[46]. Through application of a single-cell resolution analyses, we identify a distinct sub-population of *S100A8/S100A9* expressing epithelial cells. Including those markers, alongside staining for the calprotectin complex, within the epithelium, may aid with histological diagnosis of small bowel Crohn's disease, and in assessment of mucosal inflammation in known patients.

Within TN patients we also identify the upregulation of NOD-signalling genes, associated with bacterial recognition, response and proinflammatory downstream signalling[36]. Taken together, activation of these pathways infers increased inflammatory response to pathogenic bacteria, or an aberrant response to normal bacteria. Alternatively ineffectual bacteria clearance, related to a downstream 'hypoimmune' response, has recently been postulated as a cause of IBD[47]. Through WGCNA we identified a module of genes, associated with increased time to relapse, characterised by upregulation of the Th17 cell differentiation pathway. Regulation of Th17 differentiation is complex, several key cytokines, including IL1β, IL6, IL23 and TGFβ, suppressing FOXP3 expression and inducing RORC-dependant Th17 differentiation[37]. We hypothesise that in patients able to mount a robust Th17 response, bacteria are cleared quicker, resulting in immediate reduction of chronic inflammation following induction therapy. Conversely at diagnosis, most Crohn's disease patients exhibit marked upregulation of IL17- and NOD-signalling as a response to ineffective bacterial clearance, resulting in chronic inflammation. Whether IL17-signalling is driving chronic inflammation in response to invasive bacteria, or as a primary 'hyperinflammatory' response is uncertain.

However, it appears that downstream IL17-signalling within these cell populations results in production of antimicrobial peptides and proinflammatory cell infiltration as part of the disease process.

This study has several strengths. Treatment-naïve patients have no confounding effects from medications. Through a targeted sequencing approach we reduce the number of reads lost to 'house-keeping' genes, enabling accurate identification of low expression probes which impact on biological processes, a method previously successfully applied in tumour samples[13]. Additionally, this methodology provides the ability to work with low biomass samples, reduction of extraction bias and a user-friendly analytical pipeline[17]. Applying single-cell transcriptomics allowed identification of cell types driving differential gene expression and provided insight into the biology of specific cells in disease. We acknowledge the potential limitations of classifying cell types, including the specialised epithelium and monocytes, based on transcription profiles alone. However, we have followed best practice guidelines for the best-established software, SingleR, to type these cells, with confirmation from Enrichr CellType[25,35]. Nonetheless, analyses of primary patient biopsies are associated with challenges; thus, our study has some limitations. Whilst analysis was limited to Crohn's disease patients and controls there remains disease heterogeneity between individuals, exacerbated by the relatively modest numbers in each subgroup. Specifically, the effect of patient age on gene expression was unable to be assessed due to low patient numbers, however the majority of patients were aged between 10-16 years where most maturation of the immune system has already occurred[48]. Additionally, at the time of analysis, follow-up duration was insufficient to assess WGCNA and DEGs with long-term disease behaviour and response to therapy. Despite these limitations, applying novel targeted and single cell sequencing methods to Crohn's disease biopsy material, we uncover specific disease -associated pathways, identified the driver cell populations, and characterised a gene module associated with disease prognosis. Finally, we provide evidence that DropSeq sequencing in single cells and targeted assessment of expressed transcripts using HTG

EdgeSeq revealed comparable expression profiles. We restricted our comparison to the expression levels of 736 genes common to both technologies. For the sample from two patients that underwent single-cell sequencing, we pooled their cell type specific data to regenerate pseudo-bulk tissue data and compared expression levels for the 736 transcripts to those observed in the bulk tissue targeted RNASeq. Both patients independently showed highly significant (p=<0.0001) correlation of gene expression indicating reproducible results from both technologies (supplementary figure 3). This agrees with recent data from Ran et al, where targeted sequencing was validated against bulk RNA sequencing for >1200 samples[49].

Conclusion

This study demonstrates the enhanced resolution of targeted RNA sequencing to identify pathways in paediatric Crohn's disease, particularly the IL17- and NOD-signalling pathways. We identify a Th17 cell differentiation gene module associated with increased time to relapse in treatment naïve patients and utilise single-cell RNA sequencing to determine a distinct epithelial cell population driving differentially expressed genes. Personalising therapy based on underlying molecular diagnosis and stratification is an exciting prospect. Replication of these findings is required, and integration of long-term outcomes may yield improved predictive models.

Figures and tables

Figure 1- Summary of patient recruitment, sample processing and data analysis pipelines. Patients were recruited in two groups, established Crohn's disease (ED) and suspected Crohn's disease patients, consisting of treatment-naïve patients (TN) and controls. All groups underwent endoscopy with ileal biopsy. All patients had ileal biopsies retrieved and stored in RNAlater at -80°C. These biopsies underwent bulk RNA extraction and subsequent targeted RNA sequencing. A subgroup of TN patients had fresh ileal biopsies processed for single cell sequencing. Data quality control and processing steps can be seen in the figure. Integration of targeted RNA sequencing and single-cell sequencing was conducted following individual pipeline analyses.

Figure 2 - Weighted gene co-expression analysis reveals a 31-gene module specifically upregulated in treatment-naïve Crohn's disease patients. A) Normalised expression score (NES) of modules correlated with patient groups, module three genes have markedly increased expression in treatment-naïve patients only. Circle size reflects the number of genes within that module that correlate with the patient group. **B)** Mean expression of module three genes across individual patients in the three patient groups, individual lines represent individual gene expression. Each point on the X axis represents an individual patient. **C)** Identification of hub co-expression and interacting genes within module three, utilising the HitPredict database demonstrates 11 hub genes within the 31-gene module.

Figure 3- Differentially expressed genes (DEGs) were identified and characterise upregulation of antimicrobial peptides in Crohn's disease patients. A) Volcano plot demonstrating DEGs between treatment-naïve Crohn's disease (TN CD) patients vs controls. A total of 342 genes were differentially

expressed between groups, blue represents upregulated probes (n=259), red represents down regulated probes (n=83). **B)** The top 10 upregulated DEGs between TN CD patients and controls. Y axis represents log(quantile normalised) change in gene expression. Boxes represent 25-75th percentile of expression data, whiskers represent minimum and maximum data points, excluding outlier data. **C)** Volcano plot demonstrating DEGs between TN CD patients vs established Crohn's disease patients. A total of 14 genes were differentially expressed between groups, blue represents upregulated probes (n=12), red represents down regulated probes (n=2). **D)** The top 10 upregulated DEGs between TN CD patients and established Crohn's disease patients, highlighting *S100A9*, *S100A12* and *CXCL8* (IL8) as remaining significantly upregulated in TN CD patients. Y axis represents log(quantile normalised) change in gene expression. Boxes represent 25-75th percentile of expression data, whiskers represent minimum and maximum data points, excluding outlier data.

Figure 4- Hierarchical clustering of all patients using 95 genes in the NOD-signalling pathway revealed grouping of Crohn's disease patients from controls (quantile normalised data, average distance clustering). Clustering demonstrated grouping of controls together (other than 3 individuals), characterised by reduced expression of *CXCL8* (IL8), *CASP5* and *CXCL2* (cluster 3). Clusters 1 and 2, mainly consisting of Crohn's disease patients, were characterised by increased *CXCL8* (IL8) and *STAT1* expression, with cluster 2 also having increased *CXCL1* expression.

Figure 5- Single cell transcriptomics identifies monocyte and epithelial cells populations which contribute to the IL17 signature in IBD. A) UMAP plot of 1458 cells originating from two independent, digested IBD ilea samples (IBD2=710 cells, IBD3=748 cells), integrated into one neighbourhood graph using BBKNN (ScanPy, n_pcs = 50, 1999 highly variable genes (min_mean=0, max_mean=4, min_disp=0.1)). **B)** SingleR database annotation (database: BlueprintEncodeData) assigned cells into populations of CD8+ Tem, memory B-cells, monocytes, epithelial cells and plasma cells. **C)** Leiden clustering ($r = 0.5$), identified nine clusters (0-8) amongst the cell populations. **D)**

Hierarchical clustering matrix plots with top 5 marker genes (scaled UMI counts) for each Leiden cluster are displayed. **E)** Barplots displaying frequency and amplitude expression of indicated gene transcripts, that characterise treatment naïve patients in IBD. Bars are colour coded for cells as in panel **C)**, identified using Leiden clustering. Each bar shows the Scran normalised expression level of indicated transcript, in a given cell.

Table 1- Patient demographics, clinical characteristics and treatments. The two patients included in the single-cell analysis are detailed alone and are also included in treatment naïve patient column.

Acknowledgements:

We are grateful to the subjects who participated in this study. We acknowledge the use of the IRIDIS High Performance Computing Facility, together with support services at the University of Southampton.

This study was supported by the Institute for Life Sciences, University of Southampton, the National Institute for Health Research (NIHR) Southampton Biomedical Centre and the National Institute for Health Research (NIHR) Data Science Team within Southampton Biomedical Centre. The views expressed are those of the author(s) and not necessarily those of the NIHR or the Department of Health and Social Care.

We would like to thank Wessex Investigational Sciences Hub (WISH) laboratory, for providing access to WISH genomic platforms, HTG EdgeSeq and NextSeq500.

The Drop-Seq device and imaging rig utilised in this project were made by Martin Fischlechner and Jonathon West.

Funding

JJA is funded by an action medical research training fellowship and an ESPEN personal fellowship. KB is funded by Wessex Investigational Sciences Hub, Faculty of Medicine, Cancer Sciences. MEP is funded by a Sir Henry Dale Fellowship from Wellcome Trust, 109377/Z/15/Z. Drop-Seq experiments were funded by MRC grant MC_PC_15078.

References

1. Khor B., Gardet A., Xavier RJ., s GA., J. XR. Genetics and pathogenesis of inflammatory bowel disease. *Nature* 2011;**474**(7351):307–17. Doi: 10.1038/nature10209.

2. Graham DB., Xavier RJ. Pathway paradigms revealed from the genetics of inflammatory bowel disease. *Nature* 2020;**578**(7796):527–39. Doi: 10.1038/s41586-020-2025-2.

3. Kotlarz D., Beier R., Murugan D., Diestelhorst J., Jensen O., Boztug K., et al. Loss of Interleukin-10 Signaling and Infantile Inflammatory Bowel Disease: Implications for Diagnosis and Therapy. *Gastroenterology* 2012;**143**(2):347–55. Doi: 10.1053/j.gastro.2012.04.045.

4. Ashton JJ., Mossotto E., Stafford IS., Haggarty R., Coelho TAF., Batra A., et al. Genetic Sequencing of Pediatric Patients Identifies Mutations in Monogenic Inflammatory Bowel Disease Genes that Translate to Distinct Clinical Phenotypes. *Clin Transl Gastroenterol* 2020;**11**(2):e00129. Doi: 10.14309/ctg.0000000000000129.

5. Denson LA., Jurickova I., Karns R., Shaw KA., Cutler DJ., Okou DT., et al. Clinical and Genomic Correlates of Neutrophil Reactive Oxygen Species Production in Pediatric Patients With Crohn's Disease. *Gastroenterology* 2018;**154**(8):2097–110. Doi: 10.1053/j.gastro.2018.02.016.

6. Ashton JJ., Mossotto E., Ennis S., Beattie RM. Personalising medicine in inflammatory bowel disease—current and future perspectives. *Transl Pediatr* 2019;**8**(1):56. Doi: 10.21037/TP.2018.12.03.

7. West NR., Hegazy AN., Owens BMJ., Bullers SJ., Linggi B., Buonocore S., et al. Oncostatin M drives intestinal inflammation and predicts response to tumor necrosis factor–neutralizing therapy in patients with inflammatory bowel disease. *Nat Med* 2017;**23**:579–89. Doi: 10.1038/nm.4307.

8. Haberman Y., Tickle TL., Dexheimer PJ., Kim MO., Tang D., Karns R., et al. Pediatric Crohn disease patients exhibit specific ileal transcriptome and microbiome signature. *J Clin Invest* 2015;**125**(3):1363. Doi: 10.1172/JCI79657.

9. Verstockt B., Verstockt S., Dehairs J., Ballet V., Blevi H., Wollants WJ., et al. Low TREM1 expression in whole blood predicts anti-TNF response in inflammatory bowel disease. *EBioMedicine* 2019;**40**:733–42. Doi: 10.1016/j.ebiom.2019.01.027.

10. Palmer NP., Silvester JA., Lee JJ., Beam AL., Fried I., Valtchinov VI., et al. Concordance between gene expression in peripheral whole blood and colonic tissue in children with inflammatory bowel disease. *PLoS One* 2019;**14**(10):e0222952. Doi: 10.1371/journal.pone.0222952.

11. Martin DP., Miya J., Reeser JW., Roychowdhury S. Targeted RNA sequencing assay to characterize gene expression and genomic alterations. *J Vis Exp* 2016;**2016**(114). Doi: 10.3791/54090.

12. Martin-Broto J., Cruz J., Penel N., Le Cesne A., Hindi N., Luna P., et al. Pazopanib for treatment of typical solitary fibrous tumours: a multicentre, single-arm, phase 2 trial. *Lancet Oncol* 2020;**21**(3):456–66. Doi: 10.1016/s1470-2045(19)30826-5.

13. Hurtado M., Prokai L., Sankpal UT., Levesque B., Maram R., Chhabra J., et al. Next generation sequencing and functional pathway analysis to understand the mechanism of action of copper-tolfenamic acid against pancreatic cancer cells. *Process Biochem* 2020;**89**:155–64. Doi: 10.1016/j.procbio.2019.10.022.

14. Macosko EZ., Basu A., Satija R., Nemesh J., Shekhar K., Goldman M., et al. Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell* 2015;**161**(5):1202–14. Doi: 10.1016/j.cell.2015.05.002.

15. Martin JC., Chang C., Boschetti G., Ungaro R., Giri M., Grout JA., et al. Single-Cell Analysis of Crohn's Disease Lesions Identifies a Pathogenic Cellular Module Associated with Resistance to Anti-TNF Therapy. *Cell* 2019;**178**(6):1493-1508.e20. Doi: 10.1016/j.cell.2019.08.008.

16. Levine A., Koletzko S., Turner D., Escher JC., Cucchiara S., de Ridder L., et al. ESPGHAN revised porto criteria for the diagnosis of inflammatory bowel disease in children and adolescents. *J Pediatr Gastroenterol Nutr* 2014;**58**(6). Doi: 10.1097/MPG.0000000000000239.

17. HTG Autoimmune - HTG Autoimmune.

18. Godoy PM., Barczak AJ., Dehoff P., Das S., Erle DJ., Correspondence LCL. Comparison of Reproducibility, Accuracy, Sensitivity, and Specificity of miRNA Quantification Platforms n.d. Doi: 10.1016/j.celrep.2019.11.078.

19. Dillies M-A., Rau A., Aubert J., Hennequet-Antier C., Jeanmougin M., Servant N., et al. A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis. *Brief Bioinform* 2013;**14**(6):671–83. Doi: 10.1093/bib/bbs046.

20. Anders S., Huber W. Differential expression analysis for sequence count data. *Genome Biol* 2010;**11**(10):R106. Doi: 10.1186/gb-2010-11-10-r106.

21. Russo PST., Ferreira GR., Cardozo LE., Bürger MC., Arias-Carrasco R., Maruyama SR., et al. CEMiTool: A Bioconductor package for performing comprehensive modular co-expression analyses. *BMC Bioinformatics* 2018;**19**(1):56. Doi: 10.1186/s12859-018-2053-1.

22. Lopez Y., Nakai K., Patil A. HitPredict Version 4: Comprehensive Reliability Scoring of Physical Protein-Protein Interactions From More Than 100 Species - PubMed. *Database* 2015.

23. Langfelder P., Horvath S. WGCNA: An R package for weighted correlation network analysis. *BMC Bioinformatics* 2008;**9**(1):559. Doi: 10.1186/1471-2105-9-559.

24. Chen J., Bardes EE., Aronow BJ., Jegga AG. ToppGene Suite for gene list enrichment analysis and candidate gene prioritization. *Nucleic Acids Res* 2009;**37**(Web Server issue):W305-11. Doi: 10.1093/nar/gkp427.

25. Kuleshov M V., Jones MR., Rouillard AD., Fernandez NF., Duan Q., Wang Z., et al. Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. *Nucleic Acids Res* 2016;**44**(W1):W90–7. Doi: 10.1093/nar/gkw377.

26. Huang R., Grishagin I., Wang Y., Zhao T., Greene J., Obenauer JC., et al. The NCATS BioPlanet – An Integrated Platform for Exploring the Universe of Cellular Signaling Pathways for Toxicology, Systems Biology, and Chemical Genomics. *Front Pharmacol* 2019;**10**(APR):445. Doi: 10.3389/fphar.2019.00445.

27. Vallejo AF., Davies J., Grover A., Tsai C-H., Jepras R., Polak ME., et al. Resolving cellular systems by ultra-sensitive and economical single-cell transcriptome filtering. *BioRxiv* 2019:800631. Doi: 10.1101/800631.

28. Macosko EZ., Basu A., Satija R., Nemesh J., Shekhar K., Goldman M., et al. Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets. *Cell* 2015;**161**(5):1202–14. Doi: 10.1016/j.cell.2015.05.002.

29. Dobin A., Davis CA., Schlesinger F., Drenkow J., Zaleski C., Jha S., et al. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* 2013;**29**(1):15–21. Doi: 10.1093/bioinformatics/bts635.

30. Wolf FA., Angerer P., Theis FJ. SCANPY: Large-scale single-cell gene expression data analysis. *Genome Biol* 2018;**19**(1). Doi: 10.1186/s13059-017-1382-0.

31. Lun ATL., Riesenfeld S., Andrews T., Dao TP., Gomes T., Marioni JC. EmptyDrops:

Distinguishing cells from empty droplets in droplet-based single-cell RNA sequencing data. *Genome Biol* 2019;**20**(1). Doi: 10.1186/s13059-019-1662-y.

32. Lun ATL., Bach K., Marioni JC. Pooling across cells to normalize single-cell RNA sequencing data with many zero counts. *Genome Biol* 2016;**17**(1). Doi: 10.1186/s13059-016-0947-7.

33. Polański K., Young MD., Miao Z., Meyer KB., Teichmann SA., Park JE., et al. BBKNN: Fast batch alignment of single cell transcriptomes. *Bioinformatics* 2020;**36**(3):964–5. Doi: 10.1093/bioinformatics/btz625.

34. Traag VA., Waltman L., van Eck NJ. From Louvain to Leiden: guaranteeing well-connected communities. *Sci Rep* 2019;**9**(1):5233. Doi: 10.1038/s41598-019-41695-z.

35. Aran D., Looney AP., Liu L., Wu E., Fong V., Hsu A., et al. Reference-based analysis of lung single-cell sequencing reveals a transitional profibrotic macrophage. *Nat Immunol* 2019;**20**(2):163–72. Doi: 10.1038/s41590-018-0276-y.

36. Caruso R., Warner N., Inohara N., Núñez G. NOD1 and NOD2: signaling, host defense, and inflammatory disease. *Immunity* 2014;**41**(6):898–908. Doi: 10.1016/j.immuni.2014.12.010.

37. Hou G., Bishu S. Th17 Cells in Inflammatory Bowel Disease: An Update for the Clinician. *Inflamm Bowel Dis* 2020;**26**(5):653–61. Doi: 10.1093/ibd/izz316.

38. Gudjonsson JE., Ding J., Johnston A., Tejasvi T., Guzman AM., Nair RP., et al. Assessment of the psoriatic transcriptome in a large sample: Additional regulated genes and comparisons with in vitro models. *J Invest Dermatol* 2010;**130**(7):1829–40. Doi: 10.1038/jid.2010.36.

39. Gijsbers K. CXCR1-binding chemokines in inflammatory bowel diseases: down-regulated IL-8/CXCL8 production by leukocytes in Crohn's disease and selective GCP-2/CXCL6 expression in inflamed intestinal tissue. *Eur J Immunol* 2004;**34**(7):1992–2000. Doi:

10.1002/eji.200324807.

40. Uhlig HH., Schwerd T., Koletzko S., Shah N., Kammermeier J., Elkadri A., et al. The diagnostic approach to monogenic very early onset inflammatory bowel disease. *Gastroenterology* 2014;**147**(5):990-1007.e3. Doi: 10.1053/j.gastro.2014.07.023.

41. Kugathasan S., Denson LA., Walters TD., Kim M-O., Marigorta UM., Schirmer M., et al. Prediction of complicated disease course for children newly diagnosed with Crohn's disease: a multicentre inception cohort study. *Lancet* 2017;**389**(10080):1710–8. Doi: 10.1016/S0140-6736(17)30317-3.

42. Fujino S., Andoh A., Bamba S., Ogawa A., Hata K., Araki Y., et al. Increased expression of interleukin 17 in inflammatory bowel disease. *Gut* 2003;**52**(1):65–70. Doi: 10.1136/gut.52.1.65.

43. Sahin A., Calhan T., Cengiz M., Kahraman R., Aydin K., Ozdil K., et al. Serum Interleukin 17 Levels in Patients with Crohn's Disease: Real Life Data. *Dis Markers* 2014;**2014**. Doi: 10.1155/2014/690853.

44. Youssef S., Steinman L., Defea K., Jimmy DS., Lee W., Wang P., et al. IL-17 in Colonic Epithelial Cells Differential Regulation of Chemokines by. *J Immunol Ref* 2008;**181**:6536–45. Doi: 10.4049/jimmunol.181.9.6536.

45. Stawczyk-Eder K., Eder P., Lykowska-Szuber L., Krela-Kazmierczak I., Klimczak K., Szymczak A., et al. Is faecal calprotectin equally useful in all Crohn's disease locations? A prospective, comparative study. *Arch Med Sci* 2015;**11**(2):353–61. Doi: 10.5114/aoms.2014.43672.

46. Howell KJ., Kraiczy J., Nayak KM., Gasparetto M., Ross A., Lee C., et al. DNA Methylation and Transcription Patterns in Intestinal Epithelial Cells From Pediatric Patients With Inflammatory

Bowel Diseases Differentiate Disease Subtypes and Associate With Outcome. *Gastroenterology* 2018;**154**(3):585–98. Doi: 10.1053/j.gastro.2017.10.007.

47. Coelho T., Mossotto E., Gao Y., Haggarty R., Ashton JJ., Batra A., et al. Immunological Profiling of Paediatric Inflammatory Bowel Disease Using Unsupervised Machine Learning. *J Pediatr Gastroenterol Nutr* 2020:1. Doi: 10.1097/mpg.0000000000002719.

48. Simon AK., Hollander GA., McMichael A. Evolution of the immune system in humans from infancy to old age. *Proc R Soc B Biol Sci* 2015. Doi: 10.1098/rspb.2014.3085.

49. Ran D., Moharil J., Lu J., Gustafson H., Culm-Merdek K., Strand-Tibbitts K., et al. Platform comparison of HTG EdgeSeq and RNA-Seq for gene expression profiling of tumor tissue specimens. *J Clin Oncol* 2020;**38**(15_suppl):3566–3566. Doi: 10.1200/jco.2020.38.15_suppl.3566.

Table 1- Patient demographics, clinical characteristics and treatments. The two patients included in the single-cell analysis are detailed alone and are also included in treatment naïve patient column.

	Treatment naïve patients	Single cell patient SOPR 0472*	Single cell patient SOPR 0476*	Established disease patients	Controls
Number of patients	27	1	1	26	17
Number of biopsies included in analysis**	27	1	1	25	17
Mean age at diagnosis (range)	13.46 years (9.26-16.75)	13.44 years	14.81 years	12.03 years (5.79-15.3)	N/A
Mean age at ileal biopsy (range)	13.46 years (9.26-16.75)	13.44 years	14.81 years	14.64 years (8.68-17.70)	11.56 years (6.01-15.10)
Percentage female (%)	29.6% (n=8)	0% (patient is male)	0% (patient is male)	42.3% (n=11)	35.3% (n=6)
Percentage with ileitis at biopsy (histologically proven)	74% (n=20)	100% (n=1)	100% (n=1)	27% (n=7)	0%
Percentage on anti-TNF therapy	0%	0%	0%	46.2% (n=12)	0%
Percentage on thiopurine therapy	0%	0%	0%	88.5% (n=23)	0%
Percentage on Ustekinumab therapy	0%	0%	0%	7.7% (n=2)	0%

* The single-cell patients SOPR 0472 and SOPR 0476 both had ileocolonic disease (Paris classification L3) **following quality control

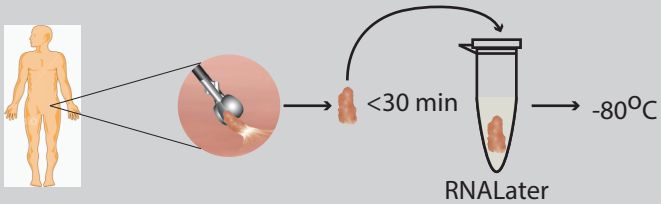
Established Crohn's disease (ED)

Suspected paediatric Crohn's disease patients

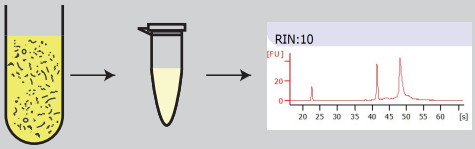
Controls

Treatment-naive Crohn's disease

Ileal biopsy, freezing into RNALater



Tissuelyser, RNA extraction , QC and quantification



Targeted sequencing of 2002 autoimmune transcripts

HTG EdgeSeq Autoimmune panel
Fully automated nuclease protection assay
Barcode adaptors and amplification
Library cleanup and normalisation
Next-generation sequencing (NextSeq500)
Single read, Read1: 50bp, Index1: 6bp, Index 2: 6bp

-Protection probes hybridise to wingmen and target RNA
-S1 nuclease digests unbound RNA and probes
-Target RNA eliminated
-PCR step adds adaptor and tags
-Quantitate pool and sequence

Targeted RNA-Seq Data analysis (bulk tissue)

bcl2fastq on IRIDIS cluster (--barcode mismatches 0),
fastq parsing (alignment, read filtering on EdgeSeq)
Analysis using HTG EdgeSeq Reveal
QC metrics (sufficent read depth >1.5M reads, minimal expression variability, sample quality/biological signal)
DGE (DESeq2), PCA (quantile normalisation)
Heatmaps and volcano plots (ward.D2)

Weighted gene co-expression network analysis (WGCNA)

CEMI tool

Hierarchical clustering

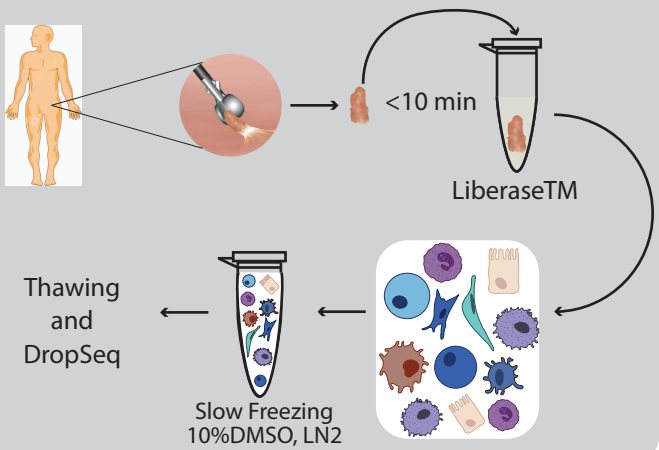
(NOD-signalling)

Gene enrichment analysis

(KEGG,BioPlanet, WIkiPathways)

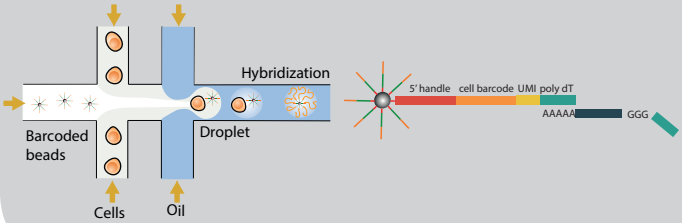
Weighted gene co-expression network analysis (WGCNA) and clinical outocme integration

Fresh Ileal biopsy, cell disaggregation, DropSeq



Single cell sequencing (DropSeq)

Digested tissue processed on ice
Coencapsulation of cells and molecular barcode beads
Cell lysis, cDNA synthesis
cDNA library tagmentation and PCR amplification
Library pooling and quantification
Next-generation sequencing (NextSeq500)
Paired end, Read1: 20bp, Read2: 50bp, Index1: 8bp



Total mRNA-Seq Data analysis (single cell)

STAR for alignment, filtering, barcode and UMI
Python based Scany with R (3.6.0), Scran normalisation
Exclusion of cells with >20% mitochondrial genes counts

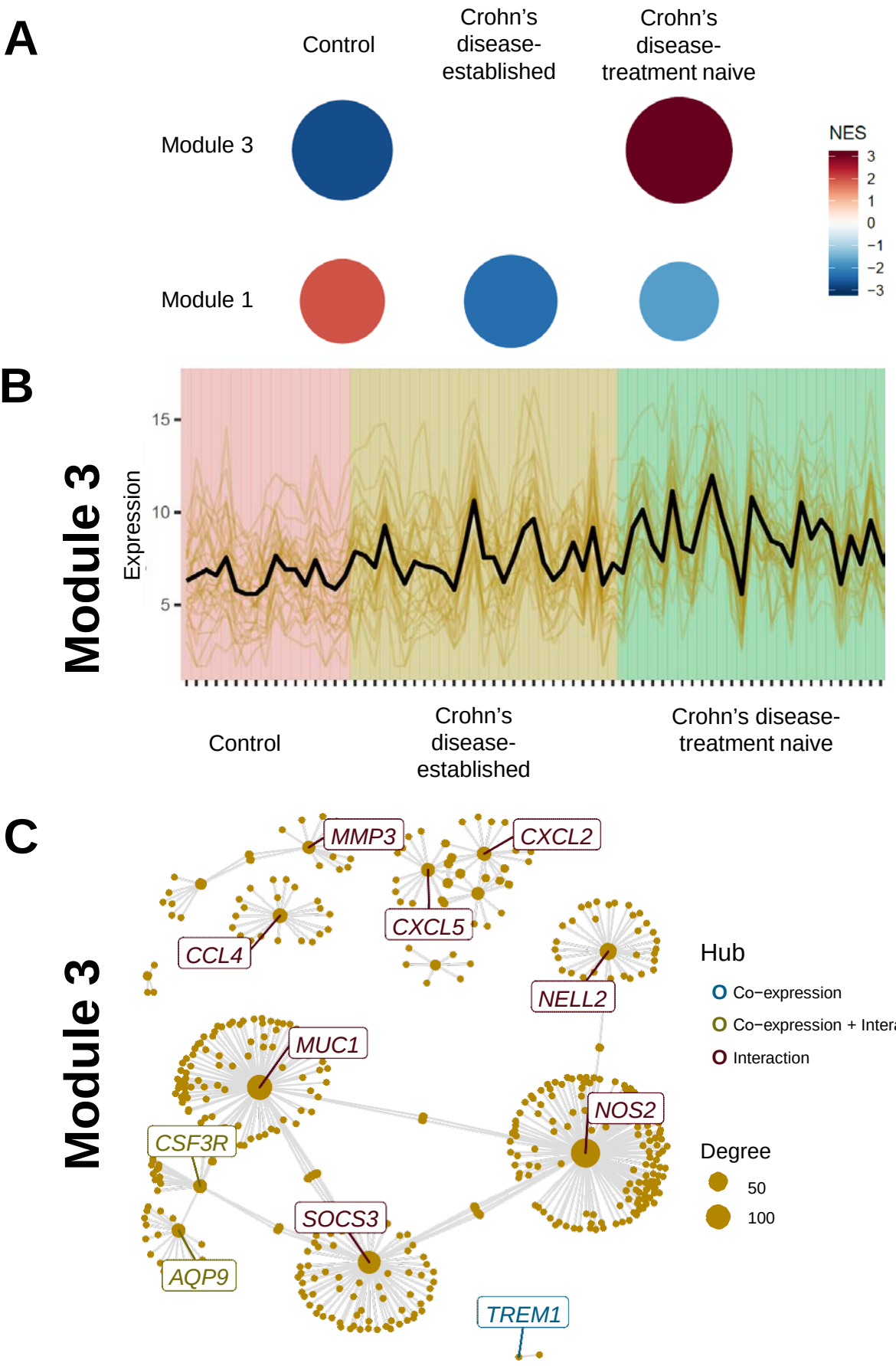
Cell neighbourhood visualisation

BBKNN-Integrated data from separate tissue samples
UMAP dimensionality reduction
Cell population annotation using SingleR
Clustering using Leiden algorithm

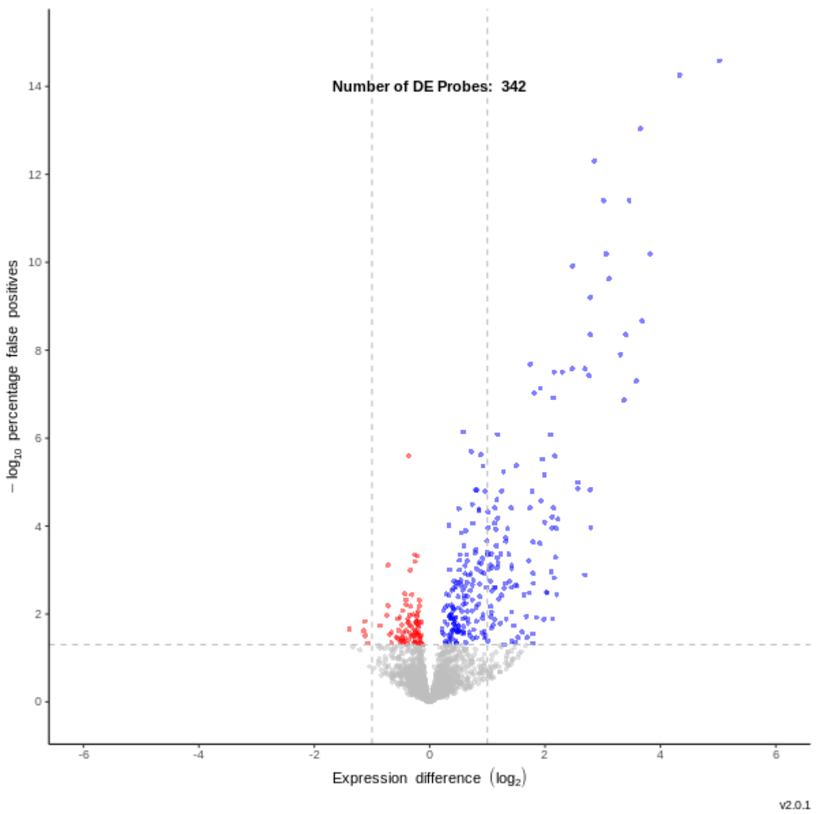
Gene enrichment analysis

Marker gene identification
ToppGene and EnRICHR

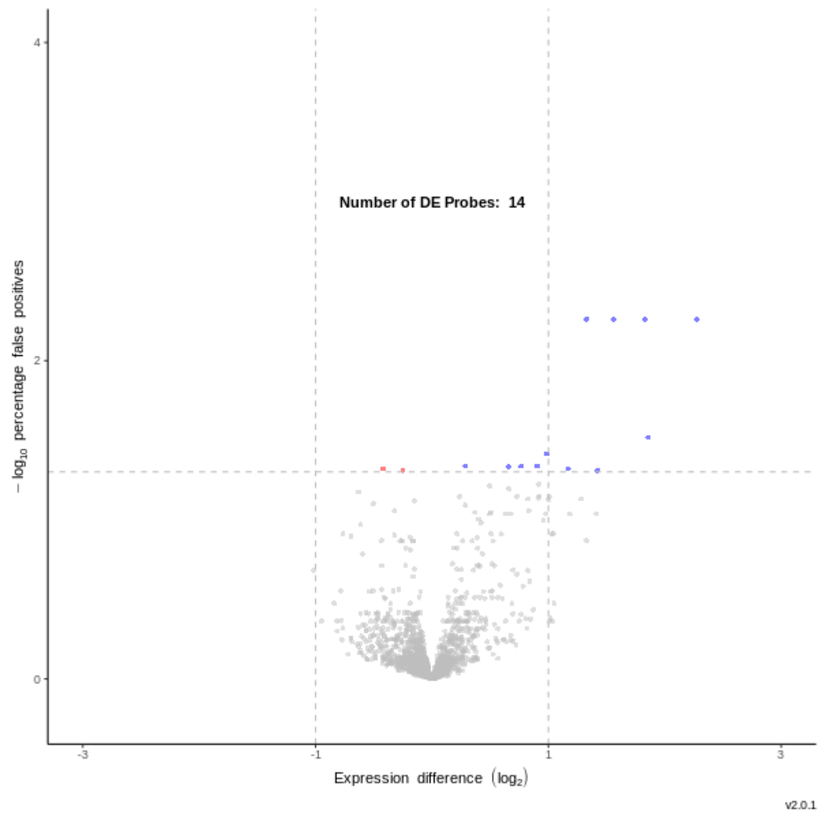
Cell types driving enriched pathways, weighted gene co-expression modules, differentially expressed genes



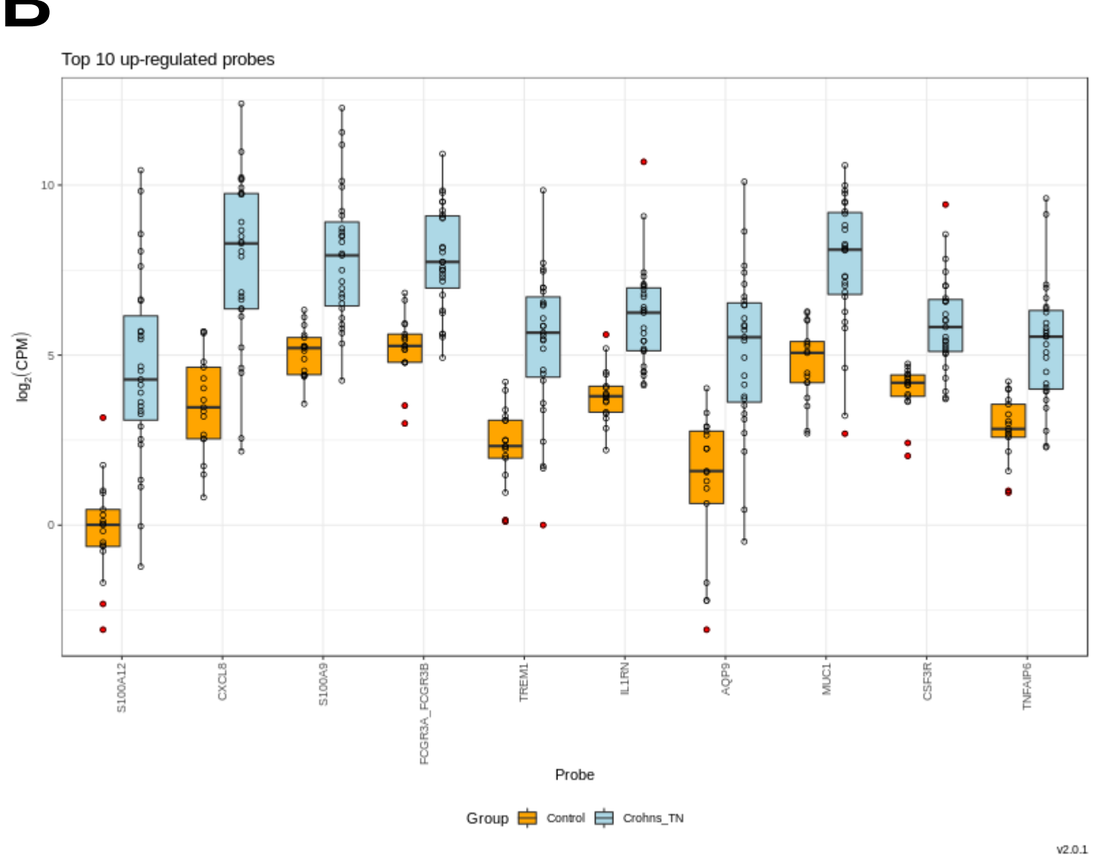
A



C



B



D