medRxiv preprint doi: https://doi.org/10.1101/2023.09.06.23295056; this version posted September 8, 2023. The copyright holder for this preprint **(which was not certified by peer review)** is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

Single-cell RNA sequencing reveals dysregulated cellular programmes in the inflamed epithelium of Crohn's disease patients.

Monika Krzak*,[1] Tobi Alegbe*,[1] D Leland Taylor*,[2] Mennatallah Ghouraba,[1] Michelle Strickland,[1] Reem Satti,[1] Tina Thompson,[3] Kenneth Arestang,[3] Moritz J Przybilla,[1] Lucia Ramirez-Navarro,[1] Bradley T Harris,[1] Kimberly Ai Xian Cheam,[4] Guillaume Noell, Steven Leonard,[1] Velislava Petrova,[5] Carla Jones-Bell,[1] Kylie R James,[6] Noor Wana,[1] May Xueqi Hu,[1] Jason Skelton,[7] Jasmin Ostermayer,[1] Yong Gu,[1] Claire Dawson,[3] Daniele Corridoni,[8,9] Cristina Cotobal Martin,[1] Miles Parkes,[3] Vivek Iyer,[1] Gareth-Rhys Jones,[10] Rebecca E. McIntyre[11], Tim Raine♀,[3] Carl A Anderson♀.[1,12]

[1] Wellcome Sanger Institute, Hinxton CB10 1SA, UK
[2] Center for Precision Health Research, National Human Genome Research Institute, Bethesda MD 20892, US
[3] Addenbrooke's Hospital, Cambridge CB2 0QQ, UK
[4] UK Dementia Research Institute, University of Cambridge, Cambridge CB2 0AH, UK
[5] Ministry of Foreign Affairs of the Republic of Bulgaria
[6] Garvan institute of medical research, Darlinghurst NSW 2010, Australia
[7] Cancer Research UK Cambridge Institute, University of Cambridge, Cambridge CB2 0RE, UK
[8] Sanofi R&D, The Bennet Building, Babraham Research Campus, Cambridge, CB22 3AT, United Kingdom.
[9] MRC Human Immunology Unit, MRC Weatherall Institute of Molecular Medicine, John Radcliffe Hospital, University of Oxford, Oxford, United Kingdom.
[10] University of Edinburgh Centre for Inflammation Research, Queens Medical Research institute, Edinburgh EH16 4TJ, UK
[11] Relation Therapeutics, London NW1 3BT, UK
[12] Lead Contact

* These authors contributed equally to this work

♀ These authors jointly supervised the study.

Correspondence: ca3@sanger.ac.uk C.A.A., tr223@cam.ac.uk T.R.

NOTE: This preprint reports new research that has not been certified by peer review and should not be used to guide clinical practice.

medRxiv preprint doi: https://doi.org/10.1101/2023.09.06.23295056; this version posted September 8, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

Summary

Crohn's disease (CD) is a complex inflammatory disorder of incompletely understood molecular aetiology. We generated a large single-cell RNA sequencing dataset from the terminal ileal biopsies of two independent cohorts comprising a total of 50 CD patients and 71 healthy controls. We performed transcriptomic analyses to reveal genes, cell types and mechanisms perturbed in CD, leveraging the power of the two cohorts to confirm our findings and assess replicability. In addition to mapping widespread alterations in cytokine signalling, we provide evidence of pan-epithelial upregulation of MHC class I genes and pathways in CD. Using non-negative matrix factorization we revealed intra- and inter-cellular upregulation of expression programmes such as G-protein coupled receptor signalling and interferon signalling, respectively, in CD. We observed an enrichment of CD heritability among marker genes for various activated T cell types and myeloid cells, supporting a causal role for these cell types in CD aetiology. Comparisons between our discovery and replication cohort revealed significant variation in differential gene-expression replicability across cell types. B, T and myeloid cells showed particularly poor replicability, suggesting caution should be exercised when interpreting unreplicated differential gene-expression results in these cell types. Overall, our results provide a rich resource for identifying cell type specific biomarkers of Crohn's disease and identifying genes, cell types and pathways that are causally and replicably associated with disease.

Keywords

Inflammatory bowel disease; IBD; single-cell RNA sequencing; Crohn's disease; differential gene expression; heritability enrichment; MHC class I; antigen-presentation

medRxiv preprint doi: https://doi.org/10.1101/2023.09.06.23295056; this version posted September 8, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

Introduction

Crohn's disease (CD) is a severely debilitating inflammatory bowel disease (IBD) characterised by chronic relapsing and remitting inflammation of the gastrointestinal tract. The biological basis of CD is poorly understood, although it is hypothesised to be caused by an overactive immune response to commensal gut bacteria in genetically predisposed individuals. The advent of therapies targeting cytokines such as TNF, IL-12 and IL-23 have improved clinical outcomes for many patients but primary non-response and secondary loss of response remains high, with 15% of CD patients requiring surgical intervention within 5 years of diagnosis (Jenkinson et al. 2020, Tsai et al. 2021). Consequently, there is an urgent need to understand more completely the aetiology of CD to broaden therapeutic hypotheses.

Genome-wide association studies (GWASs) and whole-exome sequencing (WES) studies have identified over 320 regions of the genome associated with IBD susceptibility (de Lange et al. 2017, Sazonovs et al. 2022, Liu et al. 2023). Together, these associations have suggested a broad set of immune, epithelial and stromal cell genes and pathways are involved in CD pathogenesis. However, it is challenging to establish unequivocally which specific genes and pathways are dysregulated in disease because most disease-associated variants identified to date reside in the non-coding genome. While functional genomic studies of disease-relevant tissues and cell types have shown utility for identifying CD effector genes, pathways and cell types (Soskic et al. 2019, Finucane et al. 2018, Bossini-Castillo et al. 2022, Nasrallah et al. 2020, Fairfax et al. 2014, de Lange et al. 2017), progress has been hampered by the restriction to study whole tissues or a limited number of cell types and states.

The advent of single-cell RNA sequencing (scRNA-seq) has enabled the construction of high-resolution cellular atlases of disease-relevant tissues in both health and disease. scRNA-seq studies of the gastrointestinal tract have already made some important discoveries (Kong et al. 2023), including the identification of pathogenic remodelling of mesenchymal cells (Kinchen et al. 2018) and CD8 T cells (Corridoni et al. 2020) in ulcerative colitis (UC), a separate form of IBD, as well as the identification of BEST4+ enterocytes that are crucial in maintaining luminal pH in both UC and CD (Parikh et al 2019, Smillie et al. 2019, Elmentaite et al. 2020). Comparative scRNA-seq studies in IBD have associated several gastrointestinal cell types, genes and gene expression programmes with various disease-relevant phenotypes, including drug response (Martin et al. 2019, Parikh et al 2019, Smillie et al. 2019). Replicating these early discoveries in independent datasets remains an important research goal, especially given the high level of technical noise inherent in scRNA-seq data (Jia et al. 2017) and the lack of standardised analytical approaches (Soneson and Robinson 2018, Gibson 2022).

medRxiv preprint doi: https://doi.org/10.1101/2023.09.06.23295056; this version posted September 8, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

Furthermore, most gut scRNA-seq studies to date comprise data from many thousands of cells, often taken from multiple sites in the gut, from a relatively modest number of individuals. This approach enables cellular heterogeneity within an individual to be well captured, powering the identification of rare cell types, but is relatively underpowered for the identification of disease-relevant gene expression differences between individuals. The most cost-effective means of powering differential gene expression studies using scRNA-seq is to sequence a smaller number of cells per sample, but ascertain samples from a greater number of individuals. Such designs have so far been lacking due to the high cost of single-cell RNA sequencing and the complexities associated with consenting, sampling and phenotyping a large number of IBD patients and healthy controls.

An additional complexity in the generation of scRNA-seq data from the gut is the need to process samples in a manner that permits the release of cells from the subepithelial layers. The necessary period of enzymatic digestion affects the viability and transcriptome of both epithelial cells, which undergo a process of programmed cell death when separated from the basement membrane (anoikis), and immunocytes. Here, we present a scRNA-seq data resource comprising over 440,000 cells isolated from terminal ileum biopsies from 50 CD patients with active inflammation and 71 healthy controls using a protocol designed to optimise the recovery of a wide range of gut cells and minimise artifactual changes in transcription. This study, which contains both a discovery and replication cohort, is uniquely powered to identify gastrointestinal cell types, genes and pathways associated with CD status. Using these data we identify and replicate genes that are aberrantly expressed in CD, plus those where expression is specific to given cell types and cellular processes. We then identify which of these cell types and processes are likely to play a causal role in disease by quantifying their enrichment within IBD genetic association signals. The resulting resource of terminal ileal cell transcriptional data and phenotypic information for the discovery and replication cohorts are available at zenodo DOI:10.5281/zenodo.8301000 (Data availability) and the expression atlas of gastrointestinal cell types is accessible at www.ibd-cell-portal.org.

medRxiv preprint doi: https://doi.org/10.1101/2023.09.06.23295056; this version posted September 8, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

Results

Large scale single-cell sequencing of terminal ileal biopsies identifies 49 cell clusters comprising epithelial, immune and mesenchymal cell types.

Terminal ileal (TI) biopsies were collected from 121 patients during ileo-colonoscopy, including 50 biopsies from individuals with active inflammation due to CD and 71 biopsies from healthy controls (Figure 1A). The CD patients had an average age of 39 ± 13 years, while the healthy controls had an average age of 55 ± 15 years (Table 1). The study samples were separated chronologically into two independent cohorts, with samples from the first 26 CD and 25 healthy individuals comprising the discovery cohort, and those from the subsequent 24 CD and 46 healthy controls constituting the replication cohort (Table 2).

To ensure the viability of anoikis-prone epithelial cells, which reside in close proximity to the microbiota and luminal contents of the gut, we developed a novel tissue-dissociation protocol. Biopsies were incubated with a protease that is functional at low temperatures (Adam et al. 2017), and two apoptosis inhibitors to further minimise epithelial cell death. This is followed by a brief digestion at room temperature using collagenase type IV (Methods). We have previously shown that this protocol supports the isolation of epithelial cells with higher viability and reduced cellular stress compared to conventional methods that involve collagenase digestion at 37°C (Uniken Venema et al. 2022). Additionally, the cold digest protocol better preserves the transcriptome of isolated lymphocytes.

Following the digestion of gut tissues, all cells from both cohorts were processed utilising the 10X Genomics 3' single-cell RNA sequencing platform (Methods). We identified 49 unique cell clusters from the 141,597 highest quality cells in the discovery cohort, with an average of ~3,700 CD and ~1,900 healthy post-QC cells per sample (Figure 1B). To further characterise these clusters, specifically expressed genes (SEG) were identified and manual annotation was performed based on literature references and the expression of lineage-defining markers (Table S1; Figure S1A-E; Methods). These 49 clusters delineated three major compartments, encompassing eight major cell populations: epithelial cells (enterocyte, secretory, and stem cells), immune cells (T, B, plasma B cells, and myeloid cells) and mesenchymal cells. Positional marker genes derived from a spatial transcriptomic study of mouse jejunum (Moor et al. 2018) were used to annotate enterocytes and secretory cells along the crypt-villus axis (Figure S2A). To gain further insights into cellular dynamics, we performed RNA velocity analysis (Methods), which confirmed the presence of two primary lineages of epithelial cells

medRxiv preprint doi: https://doi.org/10.1101/2023.09.06.23295056; this version posted September 8, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

originating from stem cells and differentiating into absorptive enterocytes and secretory cells along the crypt-villus axis (Figure S2B).

Gene expression signatures of all 49 cell types were used to build a cell type classifier that was subsequently applied to annotate all 187,361 QC passing cells in the discovery cohort and a further 254,551 QC passing cells in the replication cohort (with a post QC average of 5,070 cells per CD patient and 2,660 cells per healthy control). The proportion of mapped cells per cell type was highly consistent between the two cohorts (Figure S3), and the specifically expressed genes defining those cell types showed high correlation (Pearson's R > 0.77) (Figure S1F).

This comprehensive cell atlas resource was then used to quantify cell type expression specificity of 100 genes associated with monogenic forms of IBD (Figure 2A; Methods) (Bolton et al. 2022). Monogenic forms or IBD are a rare and severe group of diseases caused by highly penetrant genetic mutations in a single gene. Our atlas shows that all eight major cell type compartments of the gut harbour specific expression of at least one monogenic IBD gene, highlighting the cellular complexity of single gene disorders leading to gastrointestinal inflammation. Although previously reported as expressed on endothelial cells within the intestine (Yamaguchi et al. 2018), we show, for the first time, that *SLCO2A1*, a gene associated with Crohn's-like monogenic small intestinal ulceration, also shows highly specific expression, albeit at a lower level, on goblet and enterocyte cell populations at the top and middle of the crypt-villus axis (Figure 2B). *SLCO2A1* encodes a prostaglandin transporter that mediates the uptake and clearance of several prostaglandins, predominantly prostaglandins E2 (PGE2), E1 and E3, from various tissues. PGE2 plays a pivotal role in safeguarding the integrity of the gut epithelium against acute damage and facilitating regeneration post-injury through interactions with macrophages, epithelial cells, stromal cells, and innate lymphoid cells (Kabashima et al. 2002, Duffin et al. 2016, Roulis et al. 2014, Mukhopadhyay et al. 2020, Karaky et al. 2022). Recent work has highlighted the influence of prostaglandin E2 (PGE2) in reducing the abundance of commensal microbiota that promote regulatory T cell (Treg) function, thereby exacerbating intestinal inflammation by inhibiting mononuclear phagocytes and type I interferon signalling (Crittenden et al. 2021). Consequently, our discovery of *SLCO2A1* expression in intestinal epithelial cells located towards the top of the crypt-villus axis implicates these cells as potential contributors to inflammation in individuals harbouring homozygous loss-of-function mutations in *SLCO2A1*.

We then used our cell atlas to determine the cell type specific expression profiles of 44 strong candidate effector genes within IBD associated loci from genome-wide association studies

medRxiv preprint doi: https://doi.org/10.1101/2023.09.06.23295056; this version posted September 8, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

(Figure 2C; Methods). As observed for the monogenic IBD genes, all eight major cell populations exhibited specific expression of at least one IBD effector gene, highlighting the cellular heterogeneity and complexity of the disease. For example, *RNF186,* which controls intestinal homeostasis, and *HNF4A,* which regulates the innate immune response during inflammation, displayed highly specific gene expression in both stem cells and enterocytes. Interleukins *IL2RA*, *IL10, IL23R* and *IL18RAP* were notably more specifically expressed in CD4 T cells (Tregs or T memory cells), and expression of *NOD2* - a long established CD-associated gene, was confined to myeloid cells. Only *OSMR*, which has been previously associated with immunostimulatory effects in stromal cells (West et al. 2017), showed specific expression in mesenchymal cells. These findings collectively demonstrate the utility of our comprehensive cell atlas in enhancing the interpretation of genetic association studies of IBD by providing valuable insights into the cellular context in which disease-associated genes exert their effects.

Identification of dysregulated genes in Crohn's disease.

The unprecedented number of individuals sampled in our single-cell RNA sequencing study gives us unique power to detect gene expression differences between the inflamed biopsies from CD patients and the uninflamed biopsies from healthy controls. After comparing mean gene expression within each of the 49 cell clusters in the discovery and replication cohorts independently, we found 1,117 unique differentially expressed genes (DEGs) in the discovery cohort, and 2,406 unique DEGs in the better powered replication cohort (Figure S4A; Table S2; false discovery rate [FDR] < 5%). The number of DEGs detected for each major cell population was positively correlated ($R_{discovery}$ = 0.65, $R_{replication}$=0.97) with the number of cells, after the exclusion of plasma B cells (Figure S4B). Plasma B cells had only 92 DEGs in the discovery cohort, and 235 DEGs in the replication cohort despite the large number of sequenced cells (~47k in discovery and ~60k in replication), suggesting the transcriptome of this homogeneous population of cells is invariant with respect to disease status. Secretory epithelial cells, enterocytes and T cells had the highest number of dysregulated genes across both cohorts, with over 19-fold more DEGs than plasma B cells. We were underpowered to detect DEGs in some cell types of interest in CD, such as mesenchymal cells, due to their relative low abundance in our data (Figure S3).

Overall, 57% of significantly dysregulated genes (FDR < 5%) detected in the discovery cohort were also significantly dysregulated in the same cell type in the replication cohort. The DEG

medRxiv preprint doi: https://doi.org/10.1101/2023.09.06.23295056; this version posted September 8, 2023. The copyright holder for this preprint **(which was not certified by peer review)** is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

replication rate varied considerably by cell type, with 96% of DEGs replicating in Paneth cells but no DEGs replicating in smooth muscle cells or monocytes. Among the most up regulated genes in Crohn's disease across epithelial cell types was *PIGR*, which showed highly concordant dysregulation across both cohorts (Figure S5A). *PIGR* encodes a protein that plays a crucial role in transporting immunoglobulin A (IgA) into the intestinal lumen, where IgA shapes the intestinal microbiota through binding to a range of bacteria, which is thought to be associated with the dysbiosis of IBD (Michaud et al. 2022, Shapiro et al. 2021). Furthermore, within epithelial cells we confirmed a consistent activation of genes previously reported as being upregulated in IBD, including biomarkers *IFI27, LCN2, B2M* (Dooley et al. 2004, Yılmaz et al. 2014), and regenerating family member (REG) genes *REG1A/B* responsible for intestinal cell proliferation and differentiation (Zhang et al. 2003) (Figure S5A).

We next quantified, for each cell type, the transcriptome-wide correlation in case/control gene-expression fold changes between our discovery and replication cohort (Figure S5B). Of the eight major cell populations, four had an average Pearson's correlation coefficient (R) greater than 0.72 (stem cells, enterocytes, secretory cells, plasma B cells). The fold-change estimates for B cells, T cells, myeloid and mesenchymal cells were poorly correlated between our discovery and replication cohort ($0.15 < R < 0.51$). Within these cell types, immunoglobulin genes such as *JCHAIN, IGKC and IGHA1*, which are secreted proteins that are known to have high and variable expression, showed the poorest fold-change concordance. However, the transcriptome-wide correlation in fold-change estimates between the discovery and replication cohorts (for these cell types) remained poor even after removal of immunoglobulin genes.

To better understand the extent to which differential gene expression (DGE) effects are shared across cell types, we quantified the correlation in gene expression fold changes between cases and controls across all cell type pairs. We observed high correlation of DGE effects within all three sub-populations of stem cells (mean $R_{discovery}$ =0.91, mean $R_{replication}$ =0.93) and between stem cells and enterocyte progenitor/precursors (mean $R_{discovery}$ =0.85, mean $R_{replication}$ =0.89), suggesting that these cell types have a more homogeneous transcriptional response to disease (Figure S5C).

medRxiv preprint doi: https://doi.org/10.1101/2023.09.06.23295056; this version posted September 8, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

Dysregulation of major histocompatibility complex antigen presentation pathway in CD epithelium.

To identify biological pathways dysregulated in disease, we undertook gene set enrichment analysis across our DEG results (Methods), finding 372 and 259 dysregulated pathways (FDR < 5%) in the discovery and replication cohorts, respectively (Table S3). As expected, the per cell type replication rates of the pathway enrichment analysis mirrored those of the DEG results on which they were based. Subsequently, we chose to focus downstream analyses on the myeloid and epithelial cell populations because they exhibited the greatest number of reproducibly dysregulated pathways (n=12 unique pathways) (Figure S6A). In both cohorts, the most frequently upregulated pathway in Crohn's disease was MHC-I antigen presentation, reflecting increased expression in a suite of Human leukocyte antigen (HLA) genes, including *HLA-A, B, C* and certain non-classical MHC-I molecules such as *HLA-E* and *HLA-F* (Figure 3A-B). Although absolute expression of MHC-I family members was similar for myeloid and epithelial cells (Figure S7A), only epithelial cells upregulated MHC-I in disease. Moreover, an increased expression of MHC-I in disease was not limited to a specific subpopulation of epithelial cells but was observed across many epithelial cell types, ranging from crypt base "Enterocyte progenitors/precursors" to villus tip "Enterocyte top villus" cells, as well as goblet and Paneth cells (Figure 3B). Additionally, within epithelial cells, we noted increased expression of genes involved in the folding and peptide loading of MHC-I, including genes that coordinate peptide translocation prior to MHC-I loading (*TAP1*) and the maturation of the MHC-I complex (*CANX, CALR, TAPBP, B2M*) (Figure 3A-B). Notably, subunits of the immunoproteasome responsible for generating peptides suitable for display by MHC-I, namely *PSMB8* and *PSMB9*, were also found to be upregulated. Collectively, these consistent expression changes suggest the possibility of increased MHC-I mediated antigen presentation by epithelial cells in CD.

MHC-I genes are regulated by proximal promoters that contain binding sites for NF-kB and Interferon regulatory factor (IRF) family members, which are activated via JAK/STAT signalling pathways (Zhou 2009). Type I interferon (IFN-α/β) and type II interferon (IFN-γ) have long been implicated in the upregulation of MHC-I expression and components of its pathway (Gobin et al. 1999, Abarca-Heidemann et al. 2002, Johnson and Pober 1990) Consistent with these findings, our analysis revealed that IFN-α/β and IFN-γ signalling pathways in epithelial and myeloid cells exhibited a consistent and high degree of dysregulation in both the discovery and replication cohorts (Figure 3A). Specifically, enterocytes situated at the top and middle of the villus demonstrated a significant upregulation of the IFN-γ receptor gene, *IFNGR1* (Figure S7B). These observations suggest that IFN signalling plays a role in driving the upregulation

medRxiv preprint doi: https://doi.org/10.1101/2023.09.06.23295056; this version posted September 8, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

of MHC-I mediated antigen presentation by epithelial cells in the context of CD-associated inflammation.

Consistent with previous single-cell RNA sequencing studies of gut biopsies from CD patients, we also observed significant upregulation of MHC class II (MHC-II) genes by epithelial cells. However, this upregulation was less widespread across epithelial cells compared to MHC-I upregulation (Figure S6B). For example, 15 different epithelial cell types were significantly and replicably enriched with differentially expressed genes in the class I MHC mediated antigen processing presentation pathway, while only one cell type (a population of enterocyte progenitor cells) was significantly enriched with differentially expressed genes in the MHC class II antigen presentation pathway. MHC-II expression is classically restricted to professional antigen-presenting cells (pAPC) under the control of the *CIITA* master transactivator (Steimle et al. 1994). The restricted removal of the invariant chain (encoded by *CD74*) from the MHC-II peptide binding groove limits antigen presentation of exogenous peptides to pAPCs. Furthermore, the interaction between peptide-loaded MHC-II and CD4+ T helper cells requires the presence of co-stimulatory molecules. However, in the context of inflammation, intestinal epithelial cells have been shown to upregulate MHC-II molecules (Martin et al. 2019, Parikh et al. 2019, Maddipatla et al. 2023, Kanke et al. 2022). In our study, we predominantly observed significant upregulation of MHC-II genes *HLA-DRA, HLA-DRB1* and *HLA-DPA1* in CD enterocytes, secretory and stem cells (Figure S7C). Moreover, we observed significant upregulation of genes involved in the MHC-II antigen processing and presentation pathway, including the invariant chain (*CD74*) and cathepsin S (*CTSS*) protease, across all epithelial cell subtypes. Although we found limited evidence of upregulation of MHC-II co-stimulatory molecules on epithelial cells in patients with CD, we did detect expression of CD58 on epithelial cells, with high expression of the corresponding receptor *CD2* across T cell populations, consistent with previous reports that highlight the importance of *CD2* in the co-stimulation of intestinal T cells and regulation of inflammation (Erben et al. 2015, Pawlowski et al. 2005, Pirzer et al. 1990) (Figure S7D). Overall, our findings support an active and increased role for intestinal epithelial cells in antigen presentation to tissue-resident T cells through both MHC-I and MHC-II pathways in patients with CD.

Identification of pathogenic cellular programmes in Crohn's disease using non-negative matrix factorisation.

To more directly investigate gene expression patterns shared across cell types, we applied non-negative matrix factorisation (NMF) to our single-cell data, encompassing cells from both

medRxiv preprint doi: https://doi.org/10.1101/2023.09.06.23295056; this version posted September 8, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

CD patients and healthy controls. After optimisation, we identified 90 gene expression factors in the discovery cohort and 80 factors in the replication cohort (Figure S8A-B; Table S4; Methods). The factors detected in the two cohorts were highly correlated, with 60 of the 90 factors in the discovery cohort having at least one corresponding factor in the replication cohort (Pearson's R > 0.75) (Figure S8C). Only four factors detected in the discovery cohort showed little evidence of presence in the replication cohort (Pearson's R < 0.25). Based on the utilisation of factors across cell types, we classified them as either cell type specific or cross-cellular processes (Methods). In the discovery cohort, we identified 35 cell type-specific factors, while in the replication cohort we detected 31. Among our 49 cell types, 16 were represented by cell type-specific factors in one or both of the cohorts (Figure 4B).

The 104 remaining factors (55 in the discovery cohort and 49 in the replication cohort) represent gene expression programmes shared across cell types. These shared factors better capture the continuous nature of some gene expression across the derived discretisation of our single-cell data. The majority of the shared factors were restricted to cell types within one of our eight major cell populations. For example, a replicating factor pair (discovery factor 54, replication factor 79 [D54/R79, R = 0.9]), characterised by high expression of Treg markers (*CTLA4*, *FOXP3)* and observed in both Tregs and CD4 proliferating T cells (Figure 4A), highlights the known trajectory and plasticity of Treg/T-helper cell differentiation. We also observed several expression programmes previously described in single-cell analyses of other tissues and diseases. These include two cell-cycle factor pairs (D24/R30, R = 0.99 and D57/R15, R = 0.94) observed in proliferating cell types across the gut, a metallothionein factor pair (D58/R36, R = 0.91) used across the crypt-villus axis in epithelium and an interferon factor pair (D34/R25, R = 0.72) (Barkley et al. 2021, Kinker et al. 2020) observed across myeloid and epithelial cell subtypes.

To gain insights into the biological processes captured by the inter- and intra-cellular factors, we performed gene set enrichment analyses (GSEA; Table S5). GSEA of the cell-cycle factor pairs revealed their association with different cell-cycle stages, with D24/R30 corresponding to G1-S and D57/R15 corresponding to G2-M. These factors were heavily used in proliferating cells populations (goblet cell MKI67+, stem cell MKI67+ (1), T cell proliferating, B cell germinal centre/plasmablasts), further supporting their annotation as cell-cycle gene-expression programmes. The factor pair D34/R25, which was used across the epithelium and myeloid, showed enrichment of the same pathways (interferon response and antigen presentation) found when GSEA was performed on the epithelial cell DEGs indicating this biological signal is robust to computational approach. We fit a regression model with case/control status as a predictor (Methods) to all 190 factors and found two factors that were significantly and

medRxiv preprint doi: https://doi.org/10.1101/2023.09.06.23295056; this version posted September 8, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

replicably associated with disease (FDR < 5%, Figure 4C). A cell-specific factor, D63/R42, present in endocrine cells was positively associated with disease. D63/R42 is characterised by expression of *GIP*, a marker of K cells, and is also enriched for G-protein coupled receptor (GPCR) signalling. K cells are typically found in other regions of the small intestine such as the duodenum and the increased presence of this factor in inflamed CD terminal ileum indicates that K cells or K-like cells may play a role in dysbiosis. In another tissue, human adipocytes, increased GIP has been reported to result in increased expression of inflammatory-associated proteins such as IL-6, IL-1β and CCL2 (Pfeiffer and Keyhani-Nejad 2018) suggesting that the increased presence of this factor in disease may have an impact on inflammation. The shared B cell factor D39/R60 was negatively associated with disease, although shared across B cells it is used predominantly by the naive B cell subpopulation. The association of this factor with health potentially implies that naive B cell processes are less present in disease, which may be a result of antigen exposure inducing class switching to memory B cells and IgG-producing plasma B cells (Fleming et al. 2022) although we do not observe corresponding significant increases in corresponding factors potentially due to limited power.

Identification of disease-relevant cell types using heritability enrichment analysis

Association of expression to disease can occur either as a consequence of the disease or due to its direct contribution to pathology. To identify cell types and gene expression programmes that likely have a causal role in inflammatory bowel disease, we employed stratified linkage disequilibrium (LD) score regression (Finucane et al. 2018, Timshel et al. 2020) to integrate our single-cell RNA sequencing data with results from an IBD GWAS (de Lange et al. 2017). This analysis allowed us to test our marker genes (SEGs), differentially expressed genes (DEGs) and NMF gene expression factors for enrichment with CD and UC heritability. Furthermore, to evaluate the potential for false positives, we also quantified the extent to which these genes and factors captured the heritability of height and educational attainment, two traits where variation in gut cell gene expression is unlikely to be a contributing factor (Yengo et al. 2018, Lee et al. 2018).

We identified ten cell types with statistically significant and replicable evidence of CD heritability enrichment (family-wide error rate [FWER] < 5%) in marker genes representing ten cell types (Figure 5A). These ten cell types were situated within the myeloid population (N=4) and T cell population (N=6), which is consistent with previous studies that have partitioned IBD heritability using bulk transcriptomic data (Finucane et al. 2018). Amongst the enriched T

medRxiv preprint doi: https://doi.org/10.1101/2023.09.06.23295056; this version posted September 8, 2023. The copyright holder for this preprint **(which was not certified by peer review)** is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

cell populations, were regulatory T cells, gamma-delta T cells, CD4 memory T cells and NK cells, all of which have been shown to play a role in the pathogenesis of IBD (Mottet et al. 2003, Kelsen et al. 2011, Nemoto et al. 2013, Yusung et al. 2017), suggesting a causal role in disease that is further supported by our findings of genetic enrichment.

The two myeloid populations showing the most significant enrichment of CD heritability were monocytes and a population of cells showing characteristics intermediate between monocytes and macrophages that we annotated as "mac intermediate 2 cells". We annotated those populations based on high gene expression of monocyte-defining markers *FCGR3B, FCGR1A* and *CSF3R*, and mac intermediate 2 cells based on high expression of macrophage genes, including *FCER1G* and *TYROBP* (Dang et al. 2020), and moderate expression of macrophage-resident markers *MAF, CSF1R and C1QA/B* (Figure S1D). Monocytes also highly expressed *OSM*, a cytokine that activates stromal cells during inflammation (West et al. 2017), while mac intermediate 2 cells were the major source of tumour necrosis factor (*TNF*) expression (Figure S1D). These two cell populations were almost exclusively found in gut biopsies from CD patients, with monocytes expanded 83-fold and mac intermediate 2 cells expanded 37-fold in the discovery cohort. This significant increase in cell abundance was also observed in our replication cohort, where the two cell populations were increased 31-fold and 13-fold, respectively. Among the six cell types comprising the myeloid population, these two expanded populations showed higher expression of inflammatory markers, including *IL1A*, *IL1B*, *IL6*, *IL23A* which influence host defence and inflammasome activation (Lee et al. 2011, Seo et al 2023, Fielding et al. 2008, Sewell and Kaser 2022) and chemokines, including *CCL3/4*, which suggests they facilitate recruitment of T cells during inflammation (Honey 2006, Araujo et al. 2018). These results support existing models of CD pathogenesis based on mucosal infiltration and expansion of myeloid cell populations expressing inflammatory cytokines, accompanied by the associated infiltration and activation of pathogenic T cell populations (Baillie et al. 2017). Our study has thus provided insights into the specific cell types within these major populations that likely drive the causal mechanisms of IBD, augmenting existing models with validated gene expression data across the likely causal cell types.

We identified five cell types enriched with UC heritability, all of which were activated T cells (Figure 5A). The lower number of heritability-enriched cell types in UC compared to CD may be attributed to the reduced relevance of marker genes derived from terminal ileal biopsies from CD patients and healthy controls. In line with expectations, we did not observe any cell types enriched with heritability for our two negative control traits, height and educational

medRxiv preprint doi: https://doi.org/10.1101/2023.09.06.23295056; this version posted September 8, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

attainment. These findings underscore the importance of activated T cells in the pathogenesis of UC and highlight the specificity of these genetic enrichments to both CD and UC.

To uncover potentially causal gene expression programmes, we also tested the normalised gene expression factor z-scores for enrichment with heritability. Across the two cohorts we identified 16 factors significantly enriched in CD heritability (FWER < 5%), with 10 factors (or five factor pairs) showing significant enrichment in both cohorts (Figure 5C). Four of these five replicating factor pairs were primarily used by regulatory T cells, helper T cells, monocytes, and innate lymphoid cells ILC1/NKs, a finding that is consistent with our marker gene enrichment results and further highlights the causal role these cells play in IBD pathogenesis. The other gene expression factor with replicable evidence of heritability enrichment was one of the two factors associated with disease, D39/R60. As mentioned above D39/R60 is predominantly found in B cells which aligns with previous bulk heritability partitioning results (Finucane et al. 2018), however here we see a positive association of the genes in this factor with CD genetic signal as opposed to the prior regression which showed a negative association of cells using the factor to CD itself. Gene-set enrichment analysis unveiled several Reactome terms significantly enriched within these replicable factor pairs. Specifically, interleukin signalling was enriched in the regulatory T cell factor pair D54/R79, intercellular immunoregulatory interactions were enriched in the D70/R62 factor pair found in cytotoxic T cells and ILCs, while IFN signalling, metal sequestration and toll-like receptor (TLR) signalling were all enriched in the monocyte factor pair D30/R32. We did not find any factors that were replicably enriched with UC, height or educational attainment heritability. We also did not find any signals when we attempted to partition heritability using the differentially expressed genes (Figure 5B).

Discussion

Intestinal epithelial cells play a vital role in maintaining gut health, regulating nutrient absorption, orchestrating immune responses, and facilitating host-microbiota interactions. In the present study, we show for the first time the significant upregulation of all MHC-I genes and many of their associated molecules as a pan-epithelial response in active CD. These increases in MHC-I expression were predominately restricted to epithelial cells, where we also saw differentially expressed genes enriched within the MHC-I mediated antigen processing presentation and interferon alpha/beta and gamma signalling pathways. We also observed, albeit to a lesser extent than MHC-I, the well-established upregulation of MHC-II genes along with evidence of upregulation of related components required for antigen processing and

medRxiv preprint doi: https://doi.org/10.1101/2023.09.06.23295056; this version posted September 8, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

presentation, and intestinal CD4+ T-cell co-stimulation. To date, single-cell RNA-sequencing studies of the gut have focussed solely on the well-documented upregulation of MHC-II by intestinal epithelial cells (Hirata et al. 1986, Biton et al. 2018). Only in one of the most recent gut scRNA-seq studies has upregulation of MHC-I been seen, and this upregulation was not as widespread across epithelial cell types and MHC-I genes as observed in our data (Kong et al. 2023). Furthermore, unlike in our study, the degree of MHC-II upregulation reported in Kong et al., far exceeded that of MHC-I. While these results give us further confidence that MHC-I is indeed upregulated on the surface of epithelial cells during active CD, it is interesting to consider what factors may differentially influence the degree to which this is observed in the two datasets. Both studies compared inflamed TI biopsies from CD patients to uninflamed TI biopsies from healthy volunteers, and both comprised a mix of patients on different treatment regimens (e.g. 5-ASA, advanced therapies, steroids and immunomodulators), including medication-free patients. While the larger number of individuals in our study gives us better power to detect differentially expressed genes (50 CD patients vs 71 healthy controls in our study versus 20 CD patients and 25 controls in Kong et al. 2023), we note that the sample size of our discovery cohort (24 CD patients and 25 controls) is very similar to that of Kong et al, and we still detected more widespread differential expression of MHC-I genes in this cohort compared to MHC-II. One clear difference between the two studies is our use of our tissue digestion protocol that reduces cell stress and better inhibits anoikis compared to more traditional methods that have been applied in previous studies. Thus, we believe that our ability to detect hitherto unreported widespread alterations of MHC-I expression within the epithelium during active CD likely reflects a combination of our larger sample size and our sample processing protocol.

Recently, Heuberger and colleagues demonstrated that MHC class II antigen presentation by intestinal epithelial cells optimises bacteria-reactive CD4+ T cell responses during intestinal inflammation (Heuberger et al. 2023). However, increased epithelial MHC-I expression under inflammatory conditions is less well described and understood. Saesterstad et al. recently demonstrated, via bulk transcriptomic analysis, that a gene expression programme enriched with genes involved in MHC-I antigen processing is upregulated in the epithelium of IBD patients compared to healthy controls (Sæterstad et al. 2022). Similar alterations have been observed at the protein level through transmission electron microscopy (Bär et al. 2013), where MHC-I is predominantly localised to the basolateral membrane of intestinal epithelial cells in CD patients (Hundorfean et al. 2007). We were able to confirm and add resolution to these observations, identifying a pan-epithelial network of augmented MHC-I expression and associated molecules, associated with dysregulated interferon pathways.

medRxiv preprint doi: https://doi.org/10.1101/2023.09.06.23295056; this version posted September 8, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

Epithelial MHC-I expression driven by IFN signalling has also been reported in other tissues and disease states. Inhalation of TLR agonists can markedly increase expression of MHC-I on lung epithelial cells, a process that is dependent on type I/II/III interferon (Mathé et al. 2022). Similarly, IFN-γ-stimulated iPSC derived human intestinal organoids demonstrate increased expression of MHC-I relevant genes *PSMB9*, *HLA-F, TAP1* and *TAP2* (Workman et al. 2020). Furthermore, rotavirus, an enteric pathogen, may in part mediate its virulence by preventing MHC-I expression on epithelial cells, blocking STAT1 nuclear translocation in an IFN-dependent mechanism (Holloway et al. 2018), suggesting this pathway is key in host defence and is targeted by pathogens.

Surface expression of MHC-I can shape immune responses in both health and disease. The capacity for non-canonical antigen presentation by MHC-I on intestinal epithelial cells is unclear. Hundorfean et al. demonstrated that exogenously applied antigen was taken up and could access late endosomal compartments in human intestinal epithelial cells and colocalise with MHC-I proteins in patients with CD (Hundorfean et al. 2007). Intriguingly, this raises the possibility of “cross-presentation” of luminal antigens within the intestinal epithelium, with a display of luminal antigens to $CD8^+$ T cells. This phenomenon of cross-presentation of exogenous antigen by non-professional APCs has previously been reported in murine renal epithelium (Linke et al. 2022), and liver endothelium (Limmer et al. 2000), and in the context of the intestine suggests that epithelial cells play an active role in shaping local inflammatory responses that may be dysregulated in CD.

Taken together, our data provide evidence for upregulation of MHC-I and MHC-II in the intestinal epithelium of patients with CD, in part in an IFN-dependent manner. The upregulation of MHC-I on the epithelium may engage non-polymeric receptors on local T cells and NK cells and contribute to immune homeostasis in an antigen-independent manner, but our data also suggest the possibility that epithelial cells can present antigens to both CD4+ and CD8+ T cells and shape the inflammatory response in an antigen-specific manner. This has potential consequences for understanding the cellular contributors to the initiation and propagation of intestinal T cell responses against luminal antigens under homeostatic and inflammatory conditions.

We also found evidence of upregulation of *PIGR* in epithelial cells of patients with CD. PIGR is a basolateral transporter for dimeric immunoglobulin A (IgA) into epithelial cells, from where it undergoes a process of reverse transcytosis into the intestinal lumen as secretory IgA (sIgA). sIgA plays an important role in shaping the intestinal microbiota and binds a range of bacteria associated with the dysbiosis of IBD (Michaud et al. 2022, Shapiro et al. 2021). We, and

medRxiv preprint doi: https://doi.org/10.1101/2023.09.06.23295056; this version posted September 8, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

others, have previously shown that somatic mutations predicted to lead to a loss of function of PIGR are under positive selection in the epithelium of some patients with IBD (Olafsson et al. 2020, Kakiuchi et al. 2020), whilst *PIGR* transcription is increased in the presence of IBD-associated cytokines, including IL-17A and IFN-γ, in line with our observation of an interferon signalling signature in CD epithelial cells in our analysis of DGE. Collectively, these findings suggest that *PIGR* represents an important gene defining a cellular process involved in epithelial function, disruption of which can be associated with IBD at either a genetic or a transcriptional level and that may act through contributing to dysbiosis and loss of epithelial homeostasis.

The cell type resolution of our single-cell gut atlas, together with the large number of inflamed biopsies from CD patients and uninflamed biopsies from healthy controls, provides unprecedented power for detecting cell type specific marker genes, differentially expressed genes and programmatically expressed genes. By applying heritability enrichment analyses across these various genesets, we identified many cell types that likely play a causal role in IBD pathogenesis. Previous attempts to identify pathogenic cell types have mostly used bulk RNA sequencing data from a restricted number of cell types, tissues or states (Gettler et al. 2019, Finucane et al. 2018, Bryois et al. 2020). Despite this, these studies have implicated B cells, T cells and myeloid cells as playing a central role in the aetiology of IBD. Using our single-cell data we were able to further refine these insights to identify novel cellular subtypes and process likely driving IBD. For example, the strong enrichment of heritability among monocyte and mac intermediate 2 cell marker genes is particularly noteworthy given both populations are significantly expanded in disease in addition to highly expressing inflammatory cytokines and known IBD susceptibility genes.

There is a notable absence of genetic enrichment in non-immune cells. However, this does not lead us to conclude that these cells play no role in IBD pathogenesis. For example, our cell atlas shows that ten monogenic IBD genes are specifically expressed by non-immune cells, including *SLCO2A1*, *GUCY2C* and *IL37*. We also show that several likely effector genes within IBD associated loci are also specifically expressed by non-immune cells including *RNF186* (Beaudoin et al. 2013), *FUT2* (Franke et al. 2010), *PDLIM5* (Sazonovs et al. 2022) and *HNF4A* (UK IBD Genetics Consortium et al. 2009). Several factors could underpin the lack of heritability enrichment observed in the epithelial and stromal populations, including our exclusion of the HLA region from the analysis due to difficulties associated with partitioning heritability in regions of extended linkage disequilibrium. The underrepresentation of stromal cells in our atlas due to a combination of using pinch biopsies and digestive enzymes that

medRxiv preprint doi: https://doi.org/10.1101/2023.09.06.23295056; this version posted September 8, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

favour epithelial cell capture likely also reduces our power to detect enrichments in these cell types. We note that heritability enrichment in non-immune cells has been reported previously, for example within M cells in UC (Jagadeesh et al. 2022). While we see some evidence of UC heritability enrichment in specifically expressed genes from epithelial cells in both our discovery and replication cohort, this does not surpass our Bonferroni-corrected significance threshold.

Single-cell RNA-seq is inherently noisier than bulk RNA sequencing due to amplification bias and poorer RNA capture. This increased noise, coupled with the lack of standardised approaches to process cells and analyse single-cell data can make it challenging to draw robust insights into complex biological systems (Soneson and Robinson 2018). As a result, a great deal of effort has recently been made to compare different computational approaches for analysing single-cell data (Mou et al. 2019, Dal Molin et al. 2017, Yip et al. 2019, Krzak et al. 2019, Vieth et al. 2019) and there is a growing interest in the reproducibility of findings from single-cell sequencing studies (Gibson 2022, Skinnider et al. 2021, Vieth et al. 2019). Both to maximise and to explore the reproducibility of our findings, we generated the largest gut scRNA-seq dataset to date (by the number of individuals) and split this into a discovery and a replication cohort. All biopsies were collected from the same centre and processed using the same standardised protocols, including for tissue dissociation, sequencing data quality control (QC), clustering and auto-annotation. We observed high consistencies between the two cohorts in terms of cell type proportions and the specifically expressed genes that characterised the cell types. Our NMF and heritability enrichment analyses also showed strong evidence of replication. However, the replicability of the DGE results, and the pathway analyses that were based on these, showed marked variability between the major cell types. Power to detect differential gene expression is correlated with the number of sequenced cells. Consequently, we observed poor replicability for rarer cell types such as mesenchymal cells where the number of detected DEGs was low in both cohorts. Detecting genes that are differentially expressed in cases versus controls also requires sequencing sufficient cells from both groups. For monocytes and mac intermediate 2 cells, the rarity of these cells in the uninflamed biopsies from healthy individuals greatly reduced our power to detect DEGs. As result, the DEG replication rate was poor for these cells, despite us detecting significant heritability enrichment among specifically expressed genes within these cells (where the case/control cell ratio is not relevant). The poor replication rate with many immune cells could also reflect the greater transcription plasticity of these cells as they respond to changing environmental stimuli including diet, inflammation severity, disease stage and drug treatments. Indeed, biopsies were sampled from CD patients on a range of different treatments, including immunomodulatory and anti-cytokine therapies that might impact the transcriptome of

medRxiv preprint doi: https://doi.org/10.1101/2023.09.06.23295056; this version posted September 8, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

immunocytes in particular. Taken together, these observations provide a cautionary tale for the interpretation of scRNA-seq DGE analyses, particularly in lowly abundant or transcriptionally dynamic cells profiled in small cohorts of cases and controls.

In conclusion, we present the largest single-cell atlas of gut tissue from Crohn's disease patients and healthy individuals to date. An open-access portal for navigating our single-cell data resource has been launched at https://www.ibd-cell-portal.org/. The data available via the portal will continue to increase as we generate future releases of data. In the current study, we present a series of vignettes demonstrating how the current resource can be employed to robustly identify specifically, differentially and programmatically expressed genes and pathways in health and CD. We demonstrate how results from these analyses can be used to form biological hypotheses about CD pathogenesis for functional follow-up in downstream studies, with a particular focus on the increased expression of MHC-I genes by epithelial cells during active CD inflammation. Finally, we illustrate how these data can be combined with results from GWAS to identify candidate causal cell types and gene expression programmes underpinning IBD susceptibility.

Acknowledgements

This research was supported by the NIHR Cambridge Biomedical Research Centre (BRC-1215-20014). The views expressed are those of the authors and not necessarily those of the NIHR or the Department of Health and Social Care. This research was funded in part by the Wellcome Trust [Grant numbers 206194 and 108413/A/15/D], The Crohn's Colitis Foundation Genetics Initiative [Grant numbers 612986 and 997266] and Open Targets [OTAR2057].

We thank all individuals who kindly donated samples and their time to the study. We thank Vladimir Kiselev and Martin Prete for setting up the cellxgene interactive data explorer and Henry J Taylor for input on differential gene expression analysis.

Author contributions

Methodology, Formal Analysis and Writing – Original Draft M.K., T.A., and D.L.T;
Sample collection, T.T., K.A.,and C.D.; Sample processing, M.G., M.S., N.W., J.S., J.O., C.C.M, and M.H.; Critical discussion M.P., L.R.N, B.T.H, K.A.C., V.P., C.J., K.R.J., and M.P.; Data processing G.N., S.L., V.I., and Y.G; Writing – Review & Editing, R.S., D.C., G.R.J., and R.E.M.; Conceptualization and Supervision, C.A.A., and T.R.

medRxiv preprint doi: https://doi.org/10.1101/2023.09.06.23295056; this version posted September 8, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

Declaration of interests

C.A.A. has received consultancy or speaker fees from Genomics plc, BridgeBio and GSK. T.R. has received research/educational grants and/or speaker/consultation fees from Abbvie, Arena, Aslan, AstraZeneca, Boehringer-Ingelheim, BMS, Celgene, Ferring, Galapagos, Gilead, GSK, Heptares, LabGenius, Janssen, Mylan, MSD, Novartis, Pfizer, Sandoz, Takeda and UCB. D.C. is now an employee of Astrazeneca.

Materials and Methods

Sample ascertainment

This study was approved by the National Health Service (NHS) Research Ethics Committee (Cambridge South, REC ID 17/EE/0338). Written informed consent was given by all participants.

Individuals undergoing routine endoscopic assessment were recruited at Addenbrooke's hospital, Cambridge, UK. Clinical information and metadata for the participants are provided in Tables S1 and S2. All CD participants had a confirmed history of CD and macroscopic evidence of terminal ileal inflammation from tissue sampled during the biopsy. All control participants were undergoing endoscopic assessment or surveillance for healthy and non-cancer related reasons (e.g., history of iron deficiency anaemia, family history of colorectal cancer). Control participants did not have macroscopic evidence of intestinal inflammation, a personal history of cancer, and were not in receipt of corticosteroids or any other immune modulating therapy. Patients who were taking probiotics or antibiotics were excluded. Patients of non-European ancestries were also excluded to reduce confounding. Pinch-biopsies of the terminal ileum were collected from all participants and deposited into pre-chilled Hanks Balanced Salt Solution (HBSS) without Mg^{2+}, Ca^{2+}, or phenol red. Samples were placed on ice and immediately transferred to the Sanger Institute.

Single-cell RNA isolation and sequencing

Terminal ileal biopsies were dissociated using a single-step digestion protocol on ice to release all major intestinal cell types present in the biopsy (epithelial, immune, and stromal) without stressing the cells. First, the biopsies were mechanically minced and pipetted to release immune cells (fraction 1) from the lamina propria and the remaining tissue chunks were transferred to $HBSS^{-/-}$ containing 2 mM EDTA, 0.26 U/µl serine endoprotease isolated from *Bacillus licheniformus* (Sigma, P5380), 5 µM QVD-OPh (Abcam, ab141421), and 50 µM Y-27632 dihydrochloride (Abcam, ab120129). Tissue chunks were pipetted regularly during a

medRxiv preprint doi: https://doi.org/10.1101/2023.09.06.23295056; this version posted September 8, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

30 minute incubation on ice to release epithelial and stromal cells (fraction 2). The cells from both fractions are washed, centrifuged, and then incubated for 10 minutes at room temperature in Hank's Balanced Salt Solution (HBSS) with Mg^{2+}, Ca^{2+}, and without phenol red—including 5 mM $CaCl_2$, 1.5U/µl collagenase IV (Worthington, LS004188), and 0.1 mg/ml DNase I (Stem Cell Technologies, 07900). The cells were then filtered (30 µm; CellTrics 04-0042-2316), washed, and centrifuged before being incubated for 3 minutes at room temperature in the red blood cell lysis buffer (ACK lysis buffer; Gibco, A10492). Two final washes and centrifugations were performed before a final filtration (40 µm) and manual cell counting (haemocytometer, NanoEnTek, DHC-N01).

Single-cell RNA sequencing was undertaken using 3' 10X Genomics kits (v3.0 and v3.1) according to the manufacturer's instructions. We targeted 6,000 cells for CD participants and 3,000 cells for controls to account for the increased cellular heterogeneity in CD inflamed biopsies. Since the proportions of immune cells vary with inflammation status, we altered the ratio of cells from fractions 1 and 2 in an attempt to make the representation of cell types more equal. Viability of the mixed populations was 92±4% (mean ± S.D.) according to Trypan blue staining. Libraries were sequenced using a HiSeq4000 sequencer (Illumina; N_{CD}=24, $N_{control}$=4) or NovaSeq S4 XP sequencer (Illumina; N_{CD}=25, $N_{control}$=68) with 100bp paired-end reads, targeting 50,000 reads per cell. We compared the fraction of reads mapped confidently to the transcriptome (output metric from CellRanger) within CD and healthy participants and found no difference between sequencers (minimum p-value > 0.05, Wilcoxon rank sum test).

Single-cell RNA-seq processing and quality control procedures

CellRanger v3.0.2 was used to demultiplex reads, align reads to GRCh38 with Ensembl version 93 transcript definitions (GRCh38-3.0.0 reference file distributed by 10X Genomics), and generate cell by gene count matrices. CellBender v2.1 (Fleming et al. 2019) was then applied to identify droplets containing cells and adjust the raw counts matrix for background ambient transcript contamination (Figure S9A). For training, CellBender requires a rough estimate of the number of droplets containing cells (cell droplets) and the number of droplets without cells (empty droplets) derived from the UMI curve—the rank ordering droplet barcodes according to total UMI counts (x axis) by the total number of UMI counts per droplet (y axis). The UMI curve was calculated from droplets with a UMI count >1,000, and the threshold estimated using the "barcoderanks-inflection" procedure from DropletUtils v1.9.16 (Lun et al. 2019). To estimate the number of empty droplets, we calculated the UMI curve as described above, selected droplets with a UMI count between 250 and 10, and estimated the threshold by performing both the "barcoderanks-inflection" and "barcoderanks-knee" procedure from

medRxiv preprint doi: https://doi.org/10.1101/2023.09.06.23295056; this version posted September 8, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

DropletUtils—using 1/3rd of the distance between the two estimates as the final threshold. CellBender was run with default parameters except for excluding droplets with <10 UMI counts (--low-count-threshold) and using 300 epochs with a learning rate of $1x10^{-7}$. The final counts matrix was adjusted for the ambient transcript signature at a false positive rate of 0.1. Next, multiplets were identified and removed using scrublet v0.2.1 (Wolock et al. 2019), simulating 100,000 multiplets and calculating the multiplet threshold using the threshold_li function from the scikit-image package v0.17.2 (van der Walt et al. 2014), initialised using the threshold_otsu function. The reported sex of each sample was verified by generating pseudobulk expression matrices and comparing the expression of *XIST* to the mean expression of all genes on the Y chromosome.

De novo cell type identification

The discovery cohort (26 Crohn's disease and 25 healthy participants) was used to identify cell types and fit a model to automatically predict cell types across the entire dataset. First, additional filters were applied to ensure only the highest quality cells were used for *de novo* clustering. Cells with fewer than 100 genes expressed at ≥ 1 count, or where the percentage of counts originating from the mitochondrial genome (https://www.genenames.org/data/genegroup/#!/group/1972) was > 50, were removed. Next, an isolation forest (scikit-learn v0.23.2) was used to remove outlier cells based on (i) the percentage of counts originating from the mitochondrial genome, (ii) the total number of UMI counts per cell, (iii) the number of genes expressed (>=1 count) per cell. These metrics were selected following the recommendations of Luecken et al. (Luecken and Theis 2019).

Subsequent processing and management of the expression data was performed using scanpy v1.6.0 (Wolf et al. 2018). Genes expressed (>=1 count) in five or fewer cells across the whole dataset were removed (sc.pp.filter_genes with min_cells=5). To account for variable sequencing depth across cells, unique molecular identifier (UMI) counts were normalised by the total number of counts per cell, scaled to counts per 10,000 (CP10K; sc.pp.normalise_per_cell), and the CP10K expression matrix (ln[CP10K+1]; sc.pp.log1p) was log-transformed.

To perform dimensionality reduction, the 2,000 most variable genes across samples were selected by (i) calculating the most variable genes per sample and (ii) selecting the 2,000 genes that occurred most often across samples (sc.pp.highly_variable_genes with flavor='seurat' and batch_key=sample). After mean centering and scaling the ln(CP10K+1)

medRxiv preprint doi: https://doi.org/10.1101/2023.09.06.23295056; this version posted September 8, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

expression matrix to unit variance, principal component analysis (PCA; sc.tl.pca) was undertaken using the 2,000 most variable genes after removal of protein coding mitochondrial, ribosomal, and immunoglobulin genes, because these genes constituted the ambient signature learned by CellBender (Figure S9A). To select the number of PCs for subsequent analyses, we used a scree plot (Cattell 1966) and calculated the “knee/elbow” derived from the variance explained by each PC using the kneedle estimator v0.7.0 (Satopaa et al. 2011). From the automatically estimated elbow, we included five additional PCs in order to ensure all meaningful variability was captured, selecting 29 PCs for clustering. Finally, bbknn v1.3.12 (Polański et al. 2020) was applied to integrate samples and control for sample specific batch effects.

Clusters were defined using the Leiden graph-based clustering algorithm v0.8.3 (Traag et al. 2019) on the nearest neighbours determined by bbknn. Clusters were generated across a range of resolutions from 0.5 to 5 to empirically determine the optimal clustering resolution. For each resolution considered, the data was divided into training (2/3 of cells) and test (1/3 of cells) sets and a single layer dense neural network fit to predict cluster identity from expression using keras v2.4.3. The cluster label of each cell was predicted and the Matthews correlation coefficient (MCC) calculated for each cluster (Chicco and Jurman 2020). The final cluster classifications were chosen such that the minimum MCC across all clusters was >0.75, selecting a resolution of 3.0 (49 clusters, Figure S9B).

Cell type annotation

To determine the cell type identity of the 49 clusters, marker genes for each cluster were identified using CELLEX v1.2.1 package (Timshel et al. 2020). CELLEX calculates specifically expressed gene scores using four complementary approaches that include Gene Enrichment Score (Zeisel et al. 2018), Expression Proportion (Skene et al. 2018), Normalized Specificity Index (Dougherty et al. 2010) and Differential Expression T-statistic - the package produces a normalised mean of these four metrics which we used as our Specificity score. We used the specifically expressed genes for each cluster to label cell types through expert knowledge. To further visualise the annotated cell types, dimensionality reduction was undertaken using the uniform manifold approximation and projection (UMAP) algorithm, implemented within scanpy (scanpy.tl.umap) with default parameters, except for changing the minimum distance from 0.5 to 1.0.

medRxiv preprint doi: https://doi.org/10.1101/2023.09.06.23295056; this version posted September 8, 2023. The copyright holder for this preprint **(which was not certified by peer review)** is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

Analysing all cells that passed QC, we identified eight major cell populations including epithelial cells (stem cells, enterocytes and secretory cells), immune cells (T and B cells, plasma B cells and Myeloid cells), and mesenchymal cells.

Within epithelial cells we identified a stem cell population expressing *LGR5* and two populations expressing *MKI67*. Among enterocytes we identified three progenitor/precursor populations expressing markers *OLFM4 and KRT20.* We also identified enterocyte middle and top-villus cell types based on crypt-villus signature genes (Moor et al. 2018). Among secretory cells we identified paneth cells (*DEFA5/6*), tuft (*PLCG2*+), endocrine (*CHGA, NEUROD1*) and goblet cells at various positions of the crypt-villus axis (crypt, middle, top) based on crypt-villus signature genes *EGFR, KLF4, NT5E, SLC17A5*.

Within immune cells we identified six myeloid populations including macrophages, monocytes, and intermediate states of monocyte to macrophage differentiation expressing monocyte markers *VCAN* and *FCN1.* We found thirteen T cell populations including CD4 and CD8 T cells, ILCs and gamma-delta T cells. Within CD8 T cells we found *GZMK*+ and *FGFBP2*+ populations, earlier described in UC (Corridoni et al. 2020), showing strong enrichment in cytotoxic genes including *GZMH, KLRG1* and *NKG7*. Within CD4 T cells we identified population expressing *PASK*+ *and CCR7*+, proliferating cells (*MKI67*+), naive cells (*SELL+, LEF1*+), CD4- CD8- T cells and Treg cells (*FOXP3*+, *IL10*+) secreting *IL17A* (Hovhannisyan et al. 2011). We also found *CXCR6*+ memory CD4 T cells known for its development of the colonic inflammation by producing effector cytokines (Mandai et al. 2013) and a mix of ILC1 population with NKs (ILC1/NKs) expressing *EOMES*. Within immune cells we also found three plasma B cells IgA (*JCHAIN+, IGHA1*+) and B cells including activated (*CKS1B+, STMN1*+), naive (*IGHD+, IGHM*+), memory (*BANK1*+) and germinal centre/plasmablasts (*CD19+, TCL1A*+).

Among mesenchymal cells we found an endothelial population (*PECAM1+, VWF*+) expressing *ACKR1,* atypical chemokine receptor associated with monocyte recruitment (Pruenster et al. 2009). We also identified Fibroblasts (*COL1A1/2+, COL3A1*+), pericytes (*THY1*+) and smooth muscle cells (*ACTA2+, ACTG2*+).

We did not detect a distinct group of neutrophil and eosinophil cells - as previously well documented, they are poorly represented due to the limited ability of the 10X scRNA-seq process to capture granulocytes (Schelker et al. 2017).

medRxiv preprint doi: https://doi.org/10.1101/2023.09.06.23295056; this version posted September 8, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

Automatic cell type annotation

To label cell types across all samples, the cell type classifier trained during the *de novo* clustering process (see "*De novo* cell type identification" section) was applied to all 121 scRNA-seq samples after performing the basic processing described in the "Single-cell RNA-seq processing and quality control" section. Briefly, for each cell, a numerical vector of length 49 was generated representing the probability that a cell is each of the 49 cell types identified in the *de novo* clustering process. To keep solely high quality predictions, only those cells where the maximum probability of a given cell type was >0.5 (meaning the probability of a cell mapping to a specific cell type was greater than the sum of all other cell type probabilities) were kept. In summary, for downstream analyses we used automatically annotated cells divided into discovery (N_{CD}=25, $N_{control}$=26) and replication (N_{CD}=24, $N_{control}$=46) cohorts. These cohorts had an approximately balanced ratio between sexes (~43-47%) males (Table 2).

Crypt-villus score

We positioned epithelial cells on the crypt-villus axis and identified top villus epithelial cells and enterocyte precursors/progenitors and goblets that reside at the crypt base. We scored stem cells and enterocytes for genes such as *APOA4, APOC3, ALPI, PKIB, PMP22, SLC28A2,* derived from spatial transcriptomics that define the crypt/villus intestinal cells (Moor et al. 2018). Similarly, we used *EGFR, KLF4, NT5E, SLC17A5* signature genes to score secretory cells across the crypt/villus axis.

Velocity analysis

As the positional information of cells is largely determined by developmental history, we used RNA velocity analysis to infer the smooth transition of cells from the bottom crypt to the top villus. For velocity analysis we used scVelo v0.2.2 (Bergen et al. 2020), a kinetic model of RNA transcription and splicing. Velocity moment estimates were calculated using the scv.pp.moments function with 99 nearest neighbours and 200 PCs. Next, we computed velocities using the scv.tl.velocity function in stochastic mode, grouping by cell type with default parameters. We identified two main lineages of cells (consistent with trajectory analysis in (Smillie et al. 2019) that arise from stem cells into absorptive enterocytes and secretory cells (Figure S3), confirming well-known epithelial cell-differentiation processes.We have captured multiple villus epithelial populations, including anoikis-prone top of villus enterocytes, some of which have not previously been captured by droplet based methods. Note prior to performing velocity analysis we excluded tuft cells.

medRxiv preprint doi: https://doi.org/10.1101/2023.09.06.23295056; this version posted September 8, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

Prioritisation of IBD effector genes

Forty four genes likely to be perturbed in IBD were identified from within IBD-associated loci based on a several criteria, including but not limited to 1) presence of a coding mutation fine-mapped down to single variant resolution, 2) detailed and convincing functional follow-up work that established the causality of the gene or 3) the protein encoded by the gene plays a major role in a pathway that is targeted by an existing IBD therapy. Note, it is typically not straightforward to identify disease effector genes from within GWAS loci and this challenge remains a major focus for the field of complex disease genetics. While our list of 44 likely IBD effector genes is undoubtedly greatly enriched for true IBD effector genes, false-positives could still remain.

Differential gene expression analysis

For each cell type, we tested for association of gene expression with CD disease status using MAST v1.14.0, a two-part, generalised linear model with a logistic regression component for the discrete process (i.e., a gene is expressed or not) and linear regression component for the continuous process (i.e., the expression level) (Finak et al. 2015). For gene *i*, individual *j*, and cell *k*, let Z_{ki} indicate whether gene *i* is expressed in cell *k* and Y_{ki} denote the ln(CP10K+1) normalised gene expression. A two-part regression model was used to test for association:

$$logit(Pr(Z_{ki} = 1 \mid X_k)) = X_k\beta_i + W_k\gamma_j \quad (1)$$

$$Pr(Y_{ki} = y \mid Z_{ki} = 1) = N(X_k\beta_i + W_k\gamma_j, \sigma_i^2) \quad (2)$$

where X_k are the predictor variables for cell *k*, W_k is the random effect design matrix of cell *k* belonging to individual *j*, β_i is the vector of fixed effect regression coefficients, and γ_j is the vector of random effects (i.e., the random complement to β_i), normally distributed with mean zero and variance $\sigma^2_{\gamma k}$. CD disease status, sex, age, cell mitochondrial percentage (technical covariate associated with cellular stress), and cell complexity (i.e., the number of genes detected per cell (Finak et al. 2015, Smillie et al. 2019) were included as fixed effect variables and individual as a random effect to control for pseudoreplication bias (Zimmerman et al. 2021). The Benjamini-Hochberg procedure (Benjamini and Hochberg 1995) was used to control for multiple testing across all cell types, and p-values were obtained from the hurdle model, derived from the summed χ^2 null distributions of the discrete (Z_i) and continuous (Y_i) components, as described in Finak et al. (Finak et al. 2015). To increase the speed of each test, genes with an average CP10K of <1 in that cell type were removed prior to fitting models for each cell type.

medRxiv preprint doi: https://doi.org/10.1101/2023.09.06.23295056; this version posted September 8, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

Identifying gene expression programmes

Non-negative matrix factorisation was undertaken to derive both gene expression programmes specific to cell types and those that are shared common across multiple cell types. cNMF v1.2 (Kotliar et al. 2019) was run separately on the discovery and replication cohorts using default parameters. The optimal number of factors was estimated using the local minima of the stability metric across different numbers of factors (Figure S8A-B). Factors were defined as cell type specific if the average usage of the cell type with second highest average usage was less than 25% of the cell type with the highest average usage. Pearson's correlation was performed between all possible factor combinations in the discovery and replication cohorts with the highest R being used to define equivalent factor pairs in the discovery and replication cohort.

For each factor, we fit a linear mixed effects model to test for association with CD disease status using the R package lmerTest v3.1. Disease status, sex, age, mitochondrial percentage, and cell complexity were included as fixed effect variables (as in "*Differential gene expression analysis*" section). Individual and cell type were included as random effects to control for pseudoreplication biases.

Gene set enrichment analysis

Gene set enrichment analyses were performed using GSEA v1.17.1 (Korotkevich et al. 2016) with default parameters to identify pathways enriched among differentially expressed genes and non-negative matrix factors. Pathways were obtained from the reactome v76 gene pathway database (Fabregat et al. 2017) as part of the molecular signatures database (MSigDB) v7.4 (Subramanian et al. 2005). Z-scores from the CD vs control differential gene expression hurdle model, and the raw factor weightings from the non-negative matrix factorisation analysis, were used as input for enrichment analyses.

Heritability analysis

Heritability enrichment analysis was performed separately for the discovery and replication analyses using the CELLECT v1.3.0 workflow (Timshel et al. 2020). This workflow deploys stratified LD score regression (S-LDSC) (Finucane et al. 2018) to identify cell types with features that are enriched in genetic associations for a disease/trait of interest. CELLECT was run with default parameters, which includes filtering out complex genetic regions such as the HLA locus prior to analysis. CELLECT requires summary statistics from a genetic association study and a set of gene scores (ranging between 0 and 1) for each gene that are to be tested for heritability enrichment. For genetic summary statistics of interest, we used CD and UC statistics from (de Lange et al. 2017) and as negative controls, we used genetic summary

medRxiv preprint doi: https://doi.org/10.1101/2023.09.06.23295056; this version posted September 8, 2023. The copyright holder for this preprint **(which was not certified by peer review)** is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

statistics for height (Yengo et al. 2018) and educational attainment (Lee et al. 2018). For the gene scores, we used cell type specific gene scores and the factor gene scores from the non-negative matrix factorisation analysis. Factor gene scores were calculated by normalising (min-0.99 quantile) the positive z-scores of the factors from the non-negative matrix factorisation analysis (see “Identifying gene expression programmes"). For both types of scores (cell type specific and expression programmes) Bonferroni correction was applied to control for the total number of tests across all cell types or factors.

Data availability

Raw sequencing data files are available at the European Genome-phenome Archive (https://ega-archive.org), accession number EGAS00001003770. Processed data are available at zenodo (https://zenodo.org), accession number DOI: 10.5281/zenodo.8301000, and through https://www.ibd-cell-portal.org/.

Code availability

The code used for analyses within this study is available at https://github.com/andersonlab/sc_ti_atlas.

medRxiv preprint doi: https://doi.org/10.1101/2023.09.06.23295056; this version posted September 8, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

Supplementary Figures and Tables

Figure S1. Marker gene expression used to curate terminal ileum atlas cluster annotations.
Figure S2. Epithelial cell types represent the crypt-villus axis differentiation.
Figure S3. Cellular composition of terminal ileum atlas.
Figure S4. Differentially expressed genes in Crohn's disease (CD) across all 49 cell types.
Figure S5: Reproducibility of differentially expressed genes across discovery and replication cohorts.
Figure S6. Pathways dysregulated between health and disease.
Figure S7. Upregulation of Major Histocompatibility genes and related receptors.
Figure S8. Optimisation of non-negative matrix factorisation (NMF) parameter selection.
Figure S9. Quality control and cluster optimisation.

Table S1. Specifically expressed genes within each of the identified 49 cell types in the discovery cohort dataset.
Table S2. Significantly differentially expressed genes (FDR < 5%) within each of the identified 49 cell types in the auto-annotated discovery and replication cohort datasets.
Table S3. Significantly enriched dysregulated pathways (FDR < 5%) each of the identified 49 cell types in the auto-annotated discovery and replication cohort datasets.
Table S4. Top 20 genes by relative weighting from each non-negative matrix factor
Table S5. Top 5 significantly enriched dysregulated pathways (FDR < 5%) for each of the non-negative matrix factors

medRxiv preprint doi: https://doi.org/10.1101/2023.09.06.23295056; this version posted September 8, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

References

Bibliography

Abarca-Heidemann K, Friederichs S, Klamp T, Boehm U, Guethlein LA, Ortmann B. 2002. Regulation of the expression of mouse TAP-associated glycoprotein (tapasin) by cytokines. *Immunol Lett* **83**: 197–207.

Adam M, Potter AS, Potter SS. 2017. Psychrophilic proteases dramatically reduce single-cell RNA-seq artifacts: a molecular atlas of kidney development. *Development* **144**: 3625–3632.

Araujo JM, Gomez AC, Aguilar A, Salgado R, Balko JM, Bravo L, Doimi F, Bretel D, Morante Z, Flores C, et al. 2018. Effect of CCL5 expression in the recruitment of immune cells in triple negative breast cancer. *Sci Rep* **8**: 4899.

Baillie JK, Arner E, Daub C, De Hoon M, Itoh M, Kawaji H, Lassmann T, Carninci P, Forrest ARR, Hayashizaki Y, et al. 2017. Analysis of the human monocyte-derived macrophage transcriptome and response to lipopolysaccharide provides new insights into genetic aetiology of inflammatory bowel disease. *PLoS Genet* **13**: e1006641.

Bär F, Sina C, Hundorfean G, Pagel R, Lehnert H, Fellermann K, Büning J. 2013. Inflammatory bowel diseases influence major histocompatibility complex class I (MHC I) and II compartments in intestinal epithelial cells. *Clin Exp Immunol* **172**: 280–289.

Barkley D, Moncada R, Pour M, Liberman D, Dryg I, Werba G, Wang W, Baron M, Rao A, Xia B, et al. 2021. Recurrence of cancer cell states across diverse tumors and their interactions with the microenvironment. *BioRxiv*. doi: 10.1101/2021.12.20.473565.

Beaudoin M, Goyette P, Boucher G, Lo KS, Rivas MA, Stevens C, Alikashani A, Ladouceur M, Ellinghaus D, Törkvist L, et al. 2013. Deep resequencing of GWAS loci identifies rare variants in CARD9, IL23R and RNF186 that are associated with ulcerative colitis. *PLoS Genet* **9**: e1003723.

Benjamini Y, Hochberg Y. 1995. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society Series B (Methodological)* **57**: 289–300.

Bergen V, Lange M, Peidli S, Wolf FA, Theis FJ. 2020. Generalizing RNA velocity to transient cell states through dynamical modeling. *Nat Biotechnol* **38**: 1408–1414.

Biton M, Haber AL, Rogel N, Burgin G, Beyaz S, Schnell A, Ashenberg O, Su C-W, Smillie C, Shekhar K, et al. 2018. T helper cell cytokines modulate intestinal stem cell renewal and differentiation. *Cell* **175**: 1307-1320.e22.

Bolton C, Smillie CS, Pandey S, Elmentaite R, Wei G, Argmann C, Aschenbrenner D, James KR, McGovern DPB, Macchi M, et al. 2022. An integrated taxonomy for monogenic inflammatory bowel disease. *Gastroenterology* **162**: 859–876.

Bryois J, Skene NG, Hansen TF, Kogelman LJA, Watson HJ, Liu Z, Eating Disorders Working Group of the Psychiatric Genomics Consortium, International Headache Genetics Consortium, 23andMe Research Team, Brueggeman L, et al. 2020. Genetic identification of cell types underlying brain complex traits yields insights into the etiology

medRxiv preprint doi: https://doi.org/10.1101/2023.09.06.23295056; this version posted September 8, 2023. The copyright holder for this preprint **(which was not certified by peer review)** is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

of Parkinson's disease. *Nat Genet* **52**: 482–493.

Cattell RB. 1966. The scree test for the number of factors. *Multivariate Behav Res* **1**: 245–276.

Chicco D, Jurman G. 2020. The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genomics* **21**: 6.

Corridoni D, Antanaviciute A, Gupta T, Fawkner-Corbett D, Aulicino A, Jagielowicz M, Parikh K, Repapi E, Taylor S, Ishikawa D, et al. 2020. Single-cell atlas of colonic CD8+ T cells in ulcerative colitis. *Nat Med* **26**: 1480–1490.

Crittenden S, Goepp M, Pollock J, Robb CT, Smyth DJ, Zhou Y, Andrews R, Tyrrell V, Gkikas K, Adima A, et al. 2021. Prostaglandin E2 promotes intestinal inflammation via inhibiting microbiota-dependent regulatory T cells. *Sci Adv* **7**.

Dal Molin A, Baruzzo G, Di Camillo B. 2017. Single-Cell RNA-Sequencing: Assessment of Differential Expression Analysis Methods. *Front Genet* **8**: 62.

Dang D, Taheri S, Das S, Ghosh P, Prince LS, Sahoo D. 2020. Computational approach to identifying universal macrophage biomarkers. *Front Physiol* **11**: 275.

de Lange KM, Moutsianas L, Lee JC, Lamb CA, Luo Y, Kennedy NA, Jostins L, Rice DL, Gutierrez-Achury J, Ji S-G, et al. 2017. Genome-wide association study implicates immune activation of multiple integrin genes in inflammatory bowel disease. *Nat Genet* **49**: 256–261.

Dooley TP, Curto EV, Reddy SP, Davis RL, Lambert GW, Wilborn TW, Elson CO. 2004. Regulation of gene expression in inflammatory bowel disease and correlation with IBD drugs: screening by DNA microarrays. *Inflamm Bowel Dis* **10**: 1–14.

Dougherty JD, Schmidt EF, Nakajima M, Heintz N. 2010. Analytical approaches to RNA profiling data for the identification of genes enriched in specific cells. *Nucleic Acids Res* **38**: 4218–4230.

Duffin R, O'Connor RA, Crittenden S, Forster T, Yu C, Zheng X, Smyth D, Robb CT, Rossi F, Skouras C, et al. 2016. Prostaglandin E_2 constrains systemic inflammation through an innate lymphoid cell-IL-22 axis. *Science* **351**: 1333–1338.

Elmentaite R, Ross ADB, Roberts K, James KR, Ortmann D, Gomes T, Nayak K, Tuck L, Pritchard S, Bayraktar OA, et al. 2020. Single-Cell Sequencing of Developing Human Gut Reveals Transcriptional Links to Childhood Crohn's Disease. *Dev Cell* **55**: 771-783.e5.

Erben U, Pawlowski NN, Doerfel K, Loddenkemper C, Hoffmann JC, Siegmund B, Kühl AA. 2015. Targeting human CD2 by the monoclonal antibody CB.219 reduces intestinal inflammation in a humanized transfer colitis model. *Clin Immunol* **157**: 16–25.

Fabregat A, Sidiropoulos K, Viteri G, Forner O, Marin-Garcia P, Arnau V, D'Eustachio P, Stein L, Hermjakob H. 2017. Reactome pathway analysis: a high-performance in-memory approach. *BMC Bioinformatics* **18**: 142.

Fairfax BP, Humburg P, Makino S, Naranbhai V, Wong D, Lau E, Jostins L, Plant K, Andrews R, McGee C, et al. 2014. Innate immune activity conditions the effect of regulatory variants upon monocyte gene expression. *Science* **343**: 1246949.

Finak G, McDavid A, Yajima M, Deng J, Gersuk V, Shalek AK, Slichter CK, Miller HW,

medRxiv preprint doi: https://doi.org/10.1101/2023.09.06.23295056; this version posted September 8, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

McElrath MJ, Prlic M, et al. 2015. MAST: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data. *Genome Biol* **16**: 278.

Finucane HK, Reshef YA, Anttila V, Slowikowski K, Gusev A, Byrnes A, Gazal S, Loh P-R, Lareau C, Shoresh N, et al. 2018. Heritability enrichment of specifically expressed genes identifies disease-relevant tissues and cell types. *Nat Genet* **50**: 621–629.

Fleming A, Castro-Dopico T, Clatworthy MR. 2022. B cell class switching in intestinal immunity in health and disease. *Scand J Immunol* **95**: e13139.

Fleming SJ, Chaffin MD, Arduini A, Akkad A-D, Banks E, Marioni JC, Philippakis AA, Ellinor PT, Babadi M. 2019. Unsupervised removal of systematic background noise from droplet-based single-cell experiments using CellBender. *BioRxiv*. doi: 10.1101/791699.

Franke A, McGovern DPB, Barrett JC, Wang K, Radford-Smith GL, Ahmad T, Lees CW, Balschun T, Lee J, Roberts R, et al. 2010. Genome-wide meta-analysis increases to 71 the number of confirmed Crohn's disease susceptibility loci. *Nat Genet* **42**: 1118–1125.

Gettler K, Giri M, Kenigsberg E, Martin J, Chuang L-S, Hsu N-Y, Denson LA, Hyams JS, Griffiths A, Noe JD, et al. 2019. Prioritizing Crohn's disease genes by integrating association signals with gene expression implicates monocyte subsets. *Genes Immun* **20**: 577–588.

Gibson G. 2022. Perspectives on rigor and reproducibility in single cell genomics. *PLoS Genet* **18**: e1010210.

Gobin SJ, van Zutphen M, Woltman AM, van den Elsen PJ. 1999. Transactivation of classical and nonclassical HLA class I genes through the IFN-stimulated response element. *J Immunol* **163**: 1428–1434.

Heuberger CE, Janney A, Ilott N, Bertocchi A, Pott S, Gu Y, Pohin M, Friedrich M, Mann EH, Pearson C, et al. 2023. MHC class II antigen presentation by intestinal epithelial cells fine-tunes bacteria-reactive CD4 T cell responses. *Mucosal Immunol*. doi: 10.1016/j.mucimm.2023.05.001.

Hirata I, Austin LL, Blackwell WH, Weber JR, Dobbins WO. 1986. Immunoelectron microscopic localization of HLA-DR antigen in control small intestine and colon and in inflammatory bowel disease. *Dig Dis Sci* **31**: 1317–1330.

Holloway G, Fleming FE, Coulson BS. 2018. MHC class I expression in intestinal cells is reduced by rotavirus infection and increased in bystander cells lacking rotavirus antigen. *Sci Rep* **8**: 67.

Honey K. 2006. CCL3 and CCL4 actively recruit CD8+ T cells. *Nat Rev Immunol* **6**: 427–427.

Hovhannisyan Z, Treatman J, Littman DR, Mayer L. 2011. Characterization of interleukin-17-producing regulatory T cells in inflamed intestinal mucosa from patients with inflammatory bowel diseases. *Gastroenterology* **140**: 957–965.

Hundorfean G, Zimmer K-P, Strobel S, Gebert A, Ludwig D, Büning J. 2007. Luminal antigens access late endosomes of intestinal epithelial cells enriched in MHC I and MHC II molecules: in vivo study in Crohn's ileitis. *Am J Physiol Gastrointest Liver Physiol* **293**: G798-808.

medRxiv preprint doi: https://doi.org/10.1101/2023.09.06.23295056; this version posted September 8, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

Jenkinson PW, Plevris N, Siakavellas S, Lyons M, Arnott ID, Wilson D, Watson AJM, Jones GR, Lees CW. 2020. Temporal Trends in Surgical Resection Rates and Biologic Prescribing in Crohn's Disease: A Population-based Cohort Study. *J Crohns Colitis* **14**: 1241–1247.

Jia C, Hu Y, Kelly D, Kim J, Li M, Zhang NR. 2017. Accounting for technical noise in differential expression analysis of single-cell RNA sequencing data. *Nucleic Acids Res* **45**: 10978–10988.

Johnson DR, Pober JS. 1990. Tumor necrosis factor and immune interferon synergistically increase transcription of HLA class I heavy- and light-chain genes in vascular endothelium. *Proc Natl Acad Sci USA* **87**: 5183–5187.

Kabashima K, Saji T, Murata T, Nagamachi M, Matsuoka T, Segi E, Tsuboi K, Sugimoto Y, Kobayashi T, Miyachi Y, et al. 2002. The prostaglandin receptor EP4 suppresses colitis, mucosal damage and CD4 cell activation in the gut. *J Clin Invest* **109**: 883–893.

Kakiuchi N, Yoshida K, Uchino M, Kihara T, Akaki K, Inoue Y, Kawada K, Nagayama S, Yokoyama A, Yamamoto S, et al. 2020. Frequent mutations that converge on the NFKBIZ pathway in ulcerative colitis. *Nature* **577**: 260–265.

Kanke M, Kennedy Ng MM, Connelly S, Singh M, Schaner M, Shanahan MT, Wolber EA, Beasley C, Lian G, Jain A, et al. 2022. Single-Cell Analysis Reveals Unexpected Cellular Changes and Transposon Expression Signatures in the Colonic Epithelium of Treatment-Naïve Adult Crohn's Disease Patients. *Cell Mol Gastroenterol Hepatol* **13**: 1717–1740.

Karaky M, Boucher G, Mola S, Foisy S, Beauchamp C, Rivard M-E, Burnette M, Gosselin H, iGenoMed Consortium, Bitton A, et al. 2022. Prostaglandins and calprotectin are genetically and functionally linked to the Inflammatory Bowel Diseases. *PLoS Genet* **18**: e1010189.

Kelsen J, Dige A, Schwindt H, D'Amore F, Pedersen FS, Agnholt J, Christensen LA, Dahlerup JF, Hvas CL. 2011. Infliximab induces clonal expansion of γδ-T cells in Crohn's disease: a predictor of lymphoma risk? *PLoS ONE* **6**: e17890.

Kinchen J, Chen HH, Parikh K, Antanaviciute A, Jagielowicz M, Fawkner-Corbett D, Ashley N, Cubitt L, Mellado-Gomez E, Attar M, et al. 2018. Structural remodeling of the human colonic mesenchyme in inflammatory bowel disease. *Cell* **175**: 372-386.e17.

Kinker GS, Greenwald AC, Tal R, Orlova Z, Cuoco MS, McFarland JM, Warren A, Rodman C, Roth JA, Bender SA, et al. 2020. Pan-cancer single-cell RNA-seq identifies recurring programs of cellular heterogeneity. *Nat Genet* **52**: 1208–1218.

Kong L, Pokatayev V, Lefkovith A, Carter GT, Creasey EA, Krishna C, Subramanian S, Kochar B, Ashenberg O, Lau H, et al. 2023. The landscape of immune dysregulation in Crohn's disease revealed through single-cell transcriptomic profiling in the ileum and colon. *Immunity* **56**: 444-458.e5.

Korotkevich G, Sukhov V, Budin N, Shpak B, Artyomov MN, Sergushichev A. 2016. Fast gene set enrichment analysis. *BioRxiv*. doi: 10.1101/060012.

Kotliar D, Veres A, Nagy MA, Tabrizi S, Hodis E, Melton DA, Sabeti PC. 2019. Identifying gene expression programs of cell-type identity and cellular activity with single-cell RNA-Seq. *eLife* **8**.

medRxiv preprint doi: https://doi.org/10.1101/2023.09.06.23295056; this version posted September 8, 2023. The copyright holder for this preprint **(which was not certified by peer review)** is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

Krzak M, Raykov Y, Boukouvalas A, Cutillo L, Angelini C. 2019. Benchmark and Parameter Sensitivity Analysis of Single-Cell RNA Sequencing Clustering Methods. *Front Genet* **10**: 1253.

Lee JJ, Wedow R, Okbay A, Kong E, Maghzian O, Zacher M, Nguyen-Viet TA, Bowers P, Sidorenko J, Karlsson Linnér R, et al. 2018. Gene discovery and polygenic prediction from a genome-wide association study of educational attainment in 1.1 million individuals. *Nat Genet* **50**: 1112–1121.

Lee PY, Kumagai Y, Xu Y, Li Y, Barker T, Liu C, Sobel ES, Takeuchi O, Akira S, Satoh M, et al. 2011. IL-1α modulates neutrophil recruitment in chronic inflammation induced by hydrocarbon oil. *J Immunol* **186**: 1747–1754.

Limmer A, Ohl J, Kurts C, Ljunggren HG, Reiss Y, Groettrup M, Momburg F, Arnold B, Knolle PA. 2000. Efficient presentation of exogenous antigen by liver endothelial cells to CD8+ T cells results in antigen-specific T-cell tolerance. *Nat Med* **6**: 1348–1354.

Linke A, Cicek H, Müller A, Meyer-Schwesinger C, Melderis S, Wiech T, Wegscheid C, Ridder J, Steinmetz OM, Diehl L, et al. 2022. Antigen Cross-Presentation by Murine Proximal Tubular Epithelial Cells Induces Cytotoxic and Inflammatory CD8+ T Cells. *Cells* **11**.

Liu Z, Liu R, Gao H, Jung S, Gao X, Sun R, Liu X, Kim Y, Lee H-S, Kawai Y, et al. 2023. Genetic architecture of the inflammatory bowel diseases across East Asian and European ancestries. *Nat Genet*. doi: 10.1038/s41588-023-01384-0.

Luecken MD, Theis FJ. 2019. Current best practices in single-cell RNA-seq analysis: a tutorial. *Mol Syst Biol* **15**: e8746.

Lun ATL, Riesenfeld S, Andrews T, Dao TP, Gomes T, participants in the 1st Human Cell Atlas Jamboree, Marioni JC. 2019. EmptyDrops: distinguishing cells from empty droplets in droplet-based single-cell RNA sequencing data. *Genome Biol* **20**: 63.

Maddipatla SC, Kolachala VL, Venkateswaran S, Dodd AF, Pelia RS, Geem D, Yin H, Sun Y, Xu C, Mo A, et al. 2023. Assessing Cellular and Transcriptional Diversity of Ileal Mucosa Among Treatment-Naïve and Treated Crohn's Disease. *Inflamm Bowel Dis* **29**: 274–285.

Mandai Y, Takahashi D, Hase K, Obata Y, Furusawa Y, Ebisawa M, Nakagawa T, Sato T, Katsuno T, Saito Y, et al. 2013. Distinct Roles for CXCR6(+) and CXCR6(-) CD4(+) T Cells in the Pathogenesis of Chronic Colitis. *PLoS ONE* **8**: e65488.

Martin JC, Chang C, Boschetti G, Ungaro R, Giri M, Grout JA, Gettler K, Chuang L-S, Nayar S, Greenstein AJ, et al. 2019. Single-Cell Analysis of Crohn's Disease Lesions Identifies a Pathogenic Cellular Module Associated with Resistance to Anti-TNF Therapy. *Cell* **178**: 1493-1508.e20.

Mathé J, Benhammadi M, Kobayashi KS, Brochu S, Perreault C. 2022. Regulation of MHC Class I Expression in Lung Epithelial Cells during Inflammation. *J Immunol* **208**: 1021–1033.

McInnes et al., (2018). UMAP: Uniform Manifold Approximation and Projection. Journal of Open Source Software, 3(29), 861, https://doi.org/10.21105/joss.00861

Michaud E, Waeckel L, Gayet R, Goguyer-Deschaumes R, Chanut B, Jospin F, Bathany K,

medRxiv preprint doi: https://doi.org/10.1101/2023.09.06.23295056; this version posted September 8, 2023. The copyright holder for this preprint **(which was not certified by peer review)** is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

Monnoye M, Genet C, Prier A, et al. 2022. Alteration of microbiota antibody-mediated immune selection contributes to dysbiosis in inflammatory bowel diseases. *EMBO Mol Med* **14**: e15386.

Moor AE, Harnik Y, Ben-Moshe S, Massasa EE, Rozenberg M, Eilam R, Bahar Halpern K, Itzkovitz S. 2018. Spatial Reconstruction of Single Enterocytes Uncovers Broad Zonation along the Intestinal Villus Axis. *Cell* **175**: 1156-1167.e15.

Mottet C, Uhlig HH, Powrie F. 2003. Cutting edge: cure of colitis by CD4+CD25+ regulatory T cells. *J Immunol* **170**: 3939–3943.

Mou T, Deng W, Gu F, Pawitan Y, Vu TN. 2019. Reproducibility of Methods to Detect Differentially Expressed Genes from Single-Cell RNA Sequencing. *Front Genet* **10**: 1331.

Mukhopadhyay S, Heinz E, Porreca I, Alasoo K, Yeung A, Yang H-T, Schwerd T, Forbester JL, Hale C, Agu CA, et al. 2020. Loss of IL-10 signaling in macrophages limits bacterial killing driven by prostaglandin E2. *J Exp Med* **217**.

Nemoto Y, Kanai T, Takahara M, Oshima S, Nakamura T, Okamoto R, Tsuchiya K, Watanabe M. 2013. Bone marrow-mesenchymal stem cells are a major source of interleukin-7 and sustain colitis by forming the niche for colitogenic CD4 memory T cells. *Gut* **62**: 1142–1152.

Olafsson S, McIntyre RE, Coorens T, Butler T, Jung H, Robinson PS, Lee-Six H, Sanders MA, Arestang K, Dawson C, et al. 2020. Somatic Evolution in Non-neoplastic IBD-Affected Colon. *Cell* **182**: 672-684.e11.

Parikh K, Antanaviciute A, Fawkner-Corbett D, Jagielowicz M, Aulicino A, Lagerholm C, Davis S, Kinchen J, Chen HH, Alham NK, et al. 2019. Colonic epithelial cell diversity in health and inflammatory bowel disease. *Nature* **567**: 49–55.

Pawlowski NN, Kakirman H, Kühl AA, Liesenfeld O, Grollich K, Loddenkemper C, Zeitz M, Hoffmann JC. 2005. Alpha CD 2 mAb treatment safely attenuates adoptive transfer colitis. *Lab Invest* **85**: 1013–1023.

Pfeiffer AFH, Keyhani-Nejad F. 2018. High Glycemic Index Metabolic Damage - a Pivotal Role of GIP and GLP-1. *Trends Endocrinol Metab* **29**: 289–299.

Pirzer UC, Schürmann G, Post S, Betzler M, Meuer SC. 1990. Differential responsiveness to CD3-Ti vs. CD2-dependent activation of human intestinal T lymphocytes. *Eur J Immunol* **20**: 2339–2342.

Polański K, Young MD, Miao Z, Meyer KB, Teichmann SA, Park J-E. 2020. BBKNN: fast batch alignment of single cell transcriptomes. *Bioinformatics* **36**: 964–965.

Pruenster M, Mudde L, Bombosi P, Dimitrova S, Zsak M, Middleton J, Richmond A, Graham GJ, Segerer S, Nibbs RJB, et al. 2009. The Duffy antigen receptor for chemokines transports chemokines and supports their promigratory activity. *Nat Immunol* **10**: 101–108.

Roulis M, Nikolaou C, Kotsaki E, Kaffe E, Karagianni N, Koliaraki V, Salpea K, Ragoussis J, Aidinis V, Martini E, et al. 2014. Intestinal myofibroblast-specific Tpl2-Cox-2-PGE2 pathway links innate sensing to epithelial homeostasis. *Proc Natl Acad Sci USA* **111**: E4658-67.

medRxiv preprint doi: https://doi.org/10.1101/2023.09.06.23295056; this version posted September 8, 2023. The copyright holder for this preprint **(which was not certified by peer review)** is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

Sæterstad S, Østvik AE, Røyset ES, Bakke I, Sandvik AK, Granlund A van B. 2022. Profound gene expression changes in the epithelial monolayer of active ulcerative colitis and Crohn's disease. *PLoS ONE* **17**: e0265189.

Satopaa V, Albrecht J, Irwin D, Raghavan B. 2011. Finding a "kneedle" in a haystack: detecting knee points in system behavior. In *2011 31st International Conference on Distributed Computing Systems Workshops*, pp. 166–171, IEEE.

Sazonovs A, Stevens CR, Venkataraman GR, Yuan K, Avila B, Abreu MT, Ahmad T, Allez M, Ananthakrishnan AN, Atzmon G, et al. 2022. Large-scale sequencing identifies multiple genes and rare variants associated with Crohn's disease susceptibility. *Nat Genet* **54**: 1275–1283.

Schelker M, Feau S, Du J, Ranu N, Klipp E, MacBeath G, Schoeberl B, Raue A. 2017. Estimation of immune cell content in tumour tissue using single-cell RNA-seq data. *Nat Commun* **8**: 2032.

Shapiro JM, de Zoete MR, Palm NW, Laenen Y, Bright R, Mallette M, Bu K, Bielecka AA, Xu F, Hurtado-Lorenzo A, et al. 2021. Immunoglobulin A targets a unique subset of the microbiota in inflammatory bowel disease. *Cell Host Microbe* **29**: 83-93.e3.

Skene NG, Bryois J, Bakken TE, Breen G, Crowley JJ, Gaspar HA, Giusti-Rodriguez P, Hodge RD, Miller JA, Muñoz-Manchado AB, et al. 2018. Genetic identification of brain cell types underlying schizophrenia. *Nat Genet* **50**: 825–833.

Skinnider MA, Squair JW, Courtine G. 2021. Enabling reproducible re-analysis of single-cell data. *Genome Biol* **22**: 215.

Smillie CS, Biton M, Ordovas-Montanes J, Sullivan KM, Burgin G, Graham DB, Herbst RH, Rogel N, Slyper M, Waldman J, et al. 2019. Intra- and Inter-cellular Rewiring of the Human Colon during Ulcerative Colitis. *Cell* **178**: 714-730.e22.

Soneson C, Robinson MD. 2018. Bias, robustness and scalability in single-cell differential expression analysis. *Nat Methods* **15**: 255–261.

Steimle V, Siegrist CA, Mottet A, Lisowska-Grospierre B, Mach B. 1994. Regulation of MHC class II expression by interferon-gamma mediated by the transactivator gene CIITA. *Science* **265**: 106–109.

Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES, et al. 2005. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci USA* **102**: 15545–15550.

Timshel PN, Thompson JJ, Pers TH. 2020. Genetic mapping of etiologic brain cell types for obesity. *eLife* **9**.

Traag VA, Waltman L, van Eck NJ. 2019. From Louvain to Leiden: guaranteeing well-connected communities. *Sci Rep* **9**: 5233.

Tsai L, Ma C, Dulai PS, Prokop LJ, Eisenstein S, Ramamoorthy SL, Feagan BG, Jairath V, Sandborn WJ, Singh S. 2021. Contemporary Risk of Surgery in Patients With Ulcerative Colitis and Crohn's Disease: A Meta-Analysis of Population-Based Cohorts. *Clin Gastroenterol Hepatol* **19**: 2031-2045.e11.

UK IBD Genetics Consortium, Barrett JC, Lee JC, Lees CW, Prescott NJ, Anderson CA,

medRxiv preprint doi: https://doi.org/10.1101/2023.09.06.23295056; this version posted September 8, 2023. The copyright holder for this preprint **(which was not certified by peer review)** is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

Phillips A, Wesley E, Parnell K, Zhang H, et al. 2009. Genome-wide association study of ulcerative colitis identifies three new susceptibility loci, including the HNF4A region. *Nat Genet* **41**: 1330–1334.

Uniken Venema WTC, Ramírez-Sánchez AD, Bigaeva E, Withoff S, Jonkers I, McIntyre RE, Ghouraba M, Raine T, Weersma RK, Franke L, et al. 2022. Gut mucosa dissociation protocols influence cell type proportions and single-cell gene expression levels. *Sci Rep* **12**: 9897.

van der Walt S, Schönberger JL, Nunez-Iglesias J, Boulogne F, Warner JD, Yager N, Gouillart E, Yu T, scikit-image contributors. 2014. scikit-image: image processing in Python. *PeerJ* **2**: e453.

Vieth B, Parekh S, Ziegenhain C, Enard W, Hellmann I. 2019. A systematic evaluation of single cell RNA-seq analysis pipelines. *Nat Commun* **10**: 4667.

West NR, Hegazy AN, Owens BMJ, Bullers SJ, Linggi B, Buonocore S, Coccia M, Görtz D, This S, Stockenhuber K, et al. 2017. Oncostatin M drives intestinal inflammation and predicts response to tumor necrosis factor-neutralizing therapy in patients with inflammatory bowel disease. *Nat Med* **23**: 579–589.

Wolf FA, Angerer P, Theis FJ. 2018. SCANPY: large-scale single-cell gene expression data analysis. *Genome Biol* **19**: 15.

Wolock SL, Lopez R, Klein AM. 2019. Scrublet: Computational Identification of Cell Doublets in Single-Cell Transcriptomic Data. *Cell Syst* **8**: 281-291.e9.

Workman MJ, Troisi E, Targan SR, Svendsen CN, Barrett RJ. 2020. Modeling Intestinal Epithelial Response to Interferon-γ in Induced Pluripotent Stem Cell-Derived Human Intestinal Organoids. *Int J Mol Sci* **22**.

Wolock SL, Lopez R, Klein AM. Scrublet: Computational Identification of Cell Doublets in Single-Cell Transcriptomic Data. Cell Syst. 2019 Apr 24;8(4):281-291.e9. doi: 10.1016/j.cels.2018.11.005. Epub 2019 Apr 3. PMID: 30954476; PMCID: PMC6625319.

Yamaguchi S, Yanai S, Nakamura S, Kawasaki K, Eizuka M, Uesugi N, Sugai T, Umeno J, Esaki M, Matsumoto T. 2018. Immunohistochemical differentiation between chronic enteropathy associated with SLCO2A1 gene and other inflammatory bowel diseases. *Intest Res* **16**: 393–399.

Yengo L, Sidorenko J, Kemper KE, Zheng Z, Wood AR, Weedon MN, Frayling TM, Hirschhorn J, Yang J, Visscher PM, et al. 2018. Meta-analysis of genome-wide association studies for height and body mass index in ~700000 individuals of European ancestry. *Hum Mol Genet* **27**: 3641–3649.

Yip SH, Sham PC, Wang J. 2019. Evaluation of tools for highly variable gene discovery from single-cell RNA-seq data. *Brief Bioinformatics* **20**: 1583–1589.

Yılmaz B, Köklü S, Yüksel O, Arslan S. 2014. Serum beta 2-microglobulin as a biomarker in inflammatory bowel disease. *World J Gastroenterol* **20**: 10916–10920.

Yusung S, McGovern D, Lin L, Hommes D, Lagishetty V, Braun J. 2017. NK cells are

medRxiv preprint doi: https://doi.org/10.1101/2023.09.06.23295056; this version posted September 8, 2023. The copyright holder for this preprint **(which was not certified by peer review)** is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY 4.0 International license .

biologic and biochemical targets of 6-mercaptopurine in Crohn's disease patients. *Clin Immunol* **175**: 82–90.

Zeisel A, Hochgerner H, Lönnerberg P, Johnsson A, Memic F, van der Zwan J, Häring M, Braun E, Borm LE, La Manno G, et al. 2018. Molecular architecture of the mouse nervous system. *Cell* **174**: 999-1014.e22.

Zhang Y-W, Ding L-S, Lai M-D. 2003. Reg gene family and human diseases. *World J Gastroenterol* **9**: 2635–2641.

Zhou F. 2009. Molecular mechanisms of IFN-gamma to up-regulate MHC class I antigen processing and presentation. *Int Rev Immunol* **28**: 239–260.

Zimmerman KD, Espeland MA, Langefeld CD. 2021. A practical solution to pseudoreplication bias in single-cell studies. *Nat Commun* **12**: 738.

Figures and tables

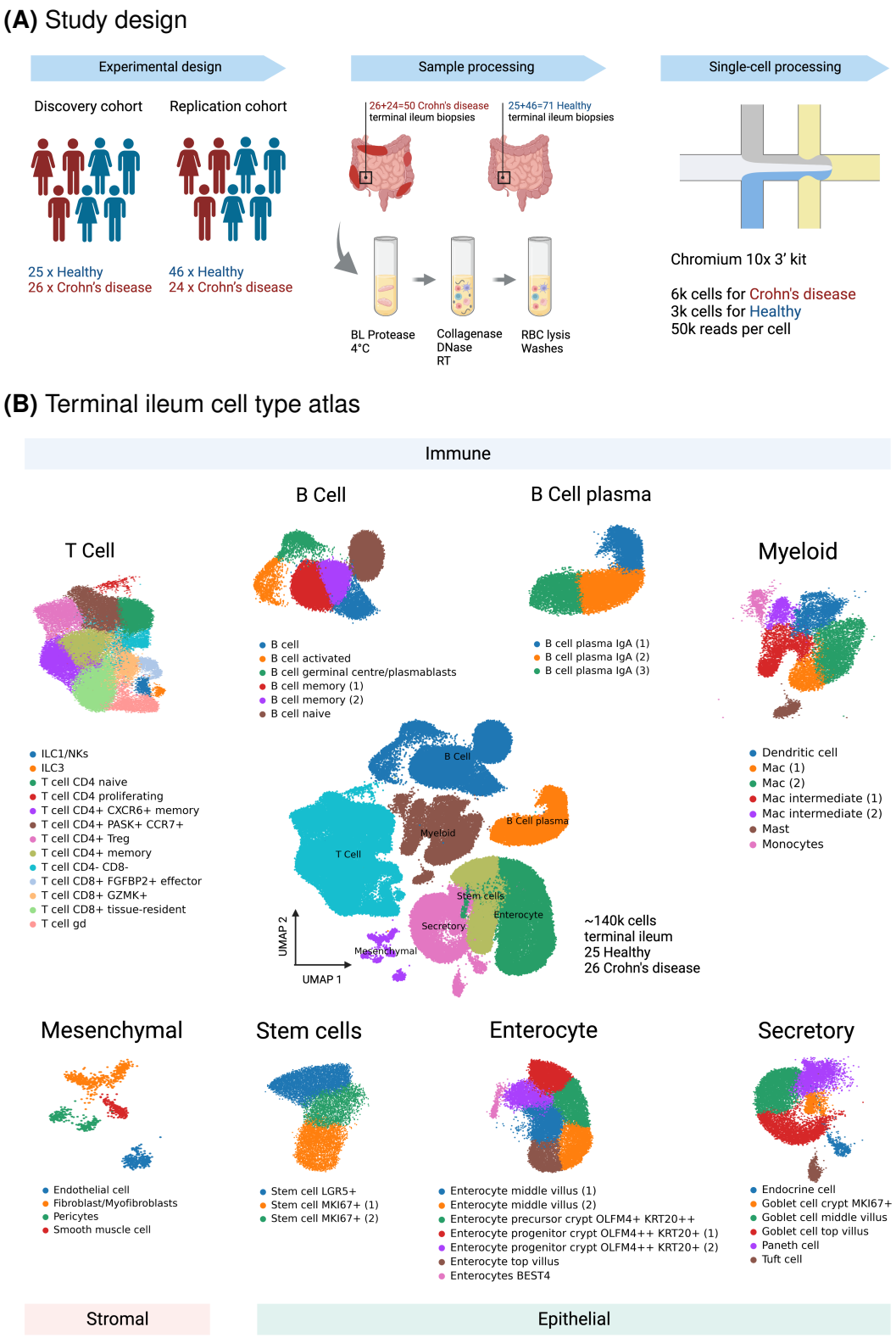
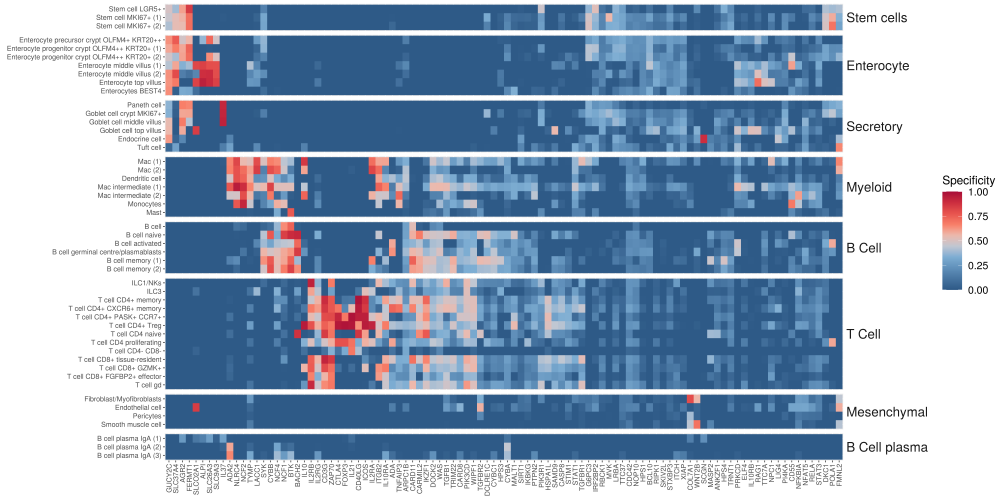

(A) Study design
Experimental design
Sample processing
Single-cell processing
Discovery cohort
Replication cohort
25 x Healthy
26 x Crohn's disease
46 x Healthy
24 x Crohn's disease
26+24=50 Crohn's disease terminal ileum biopsies
25+46=71 Healthy terminal ileum biopsies
BL Protease 4°C
Collagenase DNase RT
RBC lysis Washes
Chromium 10x 3' kit
6k cells for Crohn's disease
3k cells for Healthy
50k reads per cell
(B) Terminal ileum cell type atlas
Immune
T Cell
B Cell
B Cell plasma
Myeloid
B cell
B cell activated
B cell germinal centre/plasmablasts
B cell memory (1)
B cell memory (2)
B cell naive
B cell plasma IgA (1)
B cell plasma IgA (2)
B cell plasma IgA (3)
ILC1/NKs
ILC3
T cell CD4 naive
T cell CD4 proliferating
T cell CD4+ CXCR6+ memory
T cell CD4+ PASK+ CCR7+
T cell CD4+ Treg
T cell CD4+ memory
T cell CD4- CD8-
T cell CD8+ FGFBP2+ effector
T cell CD8+ GZMK+
T cell CD8+ tissue-resident
T cell gd
B Cell
B Cell plasma
Myeloid
T Cell
Stem cells
Enterocyte
Secretory
Mesenchymal
UMAP 2
UMAP 1
~140k cells
terminal ileum
25 Healthy
26 Crohn's disease
Dendritic cell
Mac (1)
Mac (2)
Mac intermediate (1)
Mac intermediate (2)
Mast
Monocytes
Mesenchymal
Stem cells
Enterocyte
Secretory
Endothelial cell
Fibroblast/Myofibroblasts
Pericytes
Smooth muscle cell
Stem cell LGR5+
Stem cell MKI67+ (1)
Stem cell MKI67+ (2)
Enterocyte middle villus (1)
Enterocyte middle villus (2)
Enterocyte precursor crypt OLFM4+ KRT20++
Enterocyte progenitor crypt OLFM4++ KRT20+ (1)
Enterocyte progenitor crypt OLFM4++ KRT20+ (2)
Enterocyte top villus
Enterocytes BEST4
Endocrine cell
Goblet cell crypt MKI67+
Goblet cell middle villus
Goblet cell top villus
Paneth cell
Tuft cell
Stromal
Epithelial

Figure 1

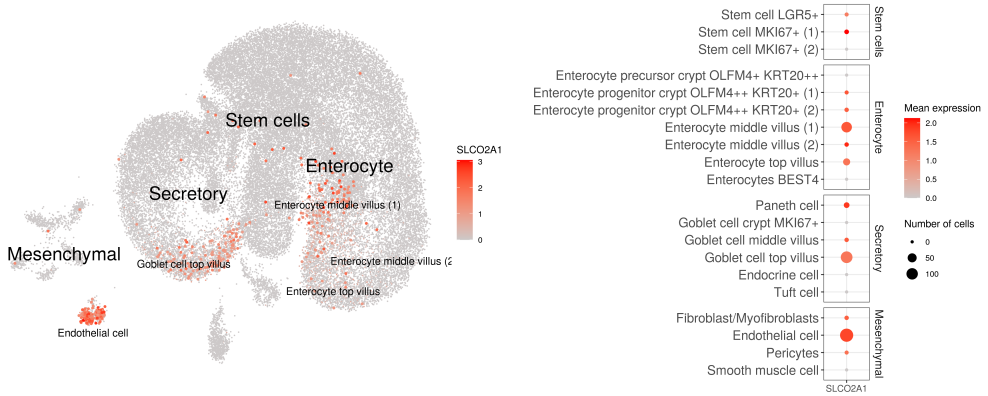
Figure 1. Single-cell expression atlas of the terminal ileum from control patients and those with active Crohn's disease.

(A) Terminal ileum biopsies from two cohorts of patients (discovery and replication) were dissociated to single cells on ice in Hank's balanced salt solution (HBSS) containing Bacillus Licheniformis (BL) protease. This was followed by a brief incubation in collagenase and then red blood cell (RBC) lysis buffer. Single cell suspensions were then profiled with the Chromium 10X 3' kit (Methods). (B) Center: UMAP projection of $\sim$140k cells passing stringent QC criteria from a discovery cohort of patients, with major cell populations represented by 8 colours. Surrounding the central UMAP, the major cell populations are further divided into cell subtypes (49 in total).

(A) Monogenic IBD genes



(B) Expression of monogenic IBD *SLCO2A1* gene across epithelial and mesenchymal cells.



(C) GWAS associated IBD genes

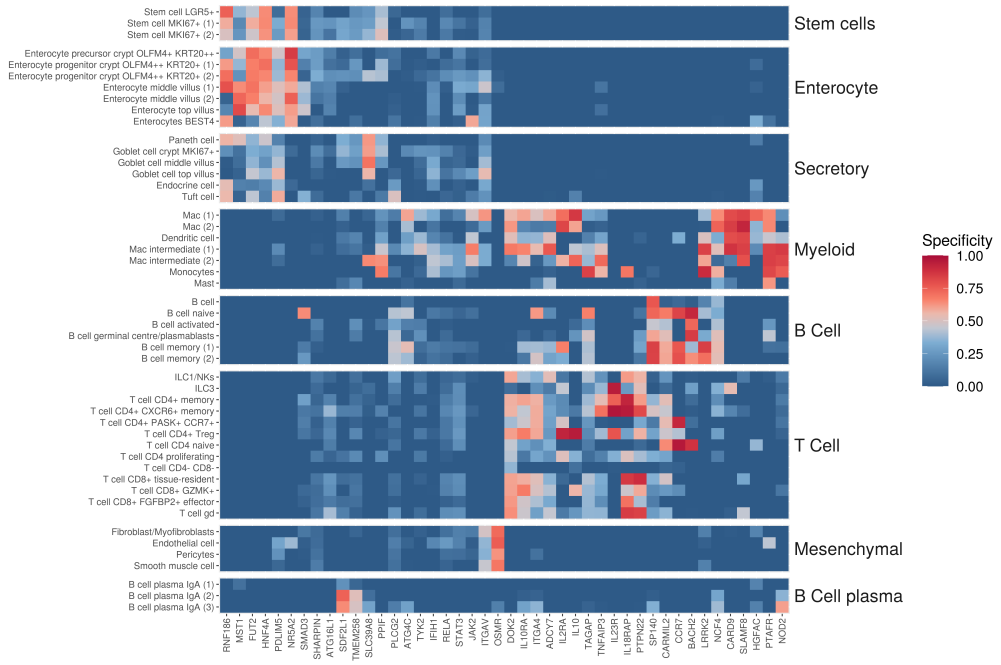
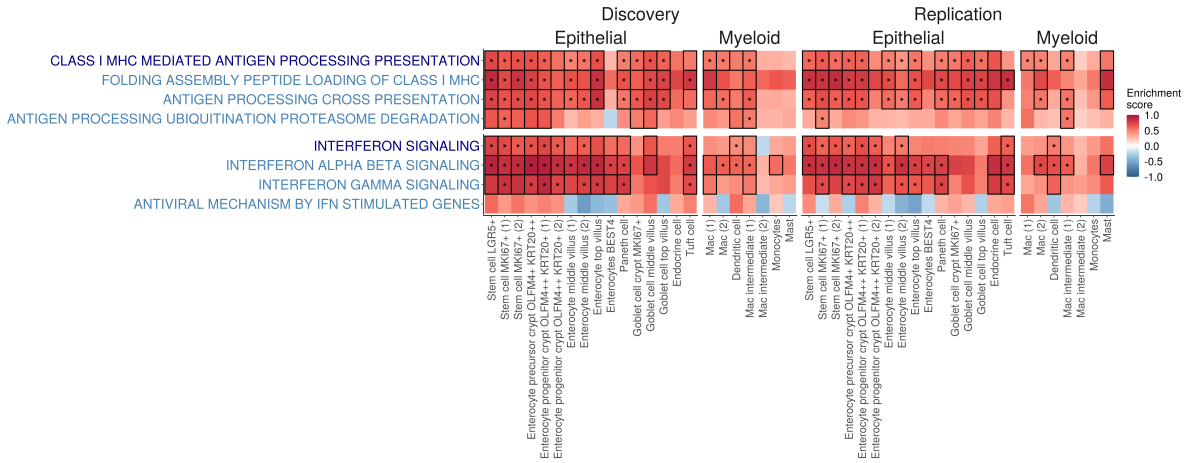


Figure 2

Figure 2. Several risk genes for inflammatory bowel disease show lineage-specific expression across cell types of the terminal ileum (TI).
(A) The monogenic IBD genes are specific to the epithelial, myeloid, B and T cell lineages (B) Terminal ileum atlas UMAP projection (from discovery dataset only) was subset to epithelial and mesenchymal cells and coloured by the log1p CP10k expression of *SLCO2A1* monogenic IBD gene. Alongside dotplot representing the number of cells expressing *SLCO2A1* with log1p CP10k > 1. (C) Genes identified through genome wide association studies (GWAS) of IBD were prioritised using several criteria (Methods) and shown to be specifically expressed across epithelial, myeloid, B and T cell lineages. Specificity displayed is calculated as a mean of discovery and replication gene specificity scores (Methods) and represents how specific gene expression is to each of identified 49 cell types in the terminal ileum.

(A) Enrichment of epithelial and myeloid cells in immune pathways



(B) Dysregulated genes in class MHC I mediated antigen presentation pathway

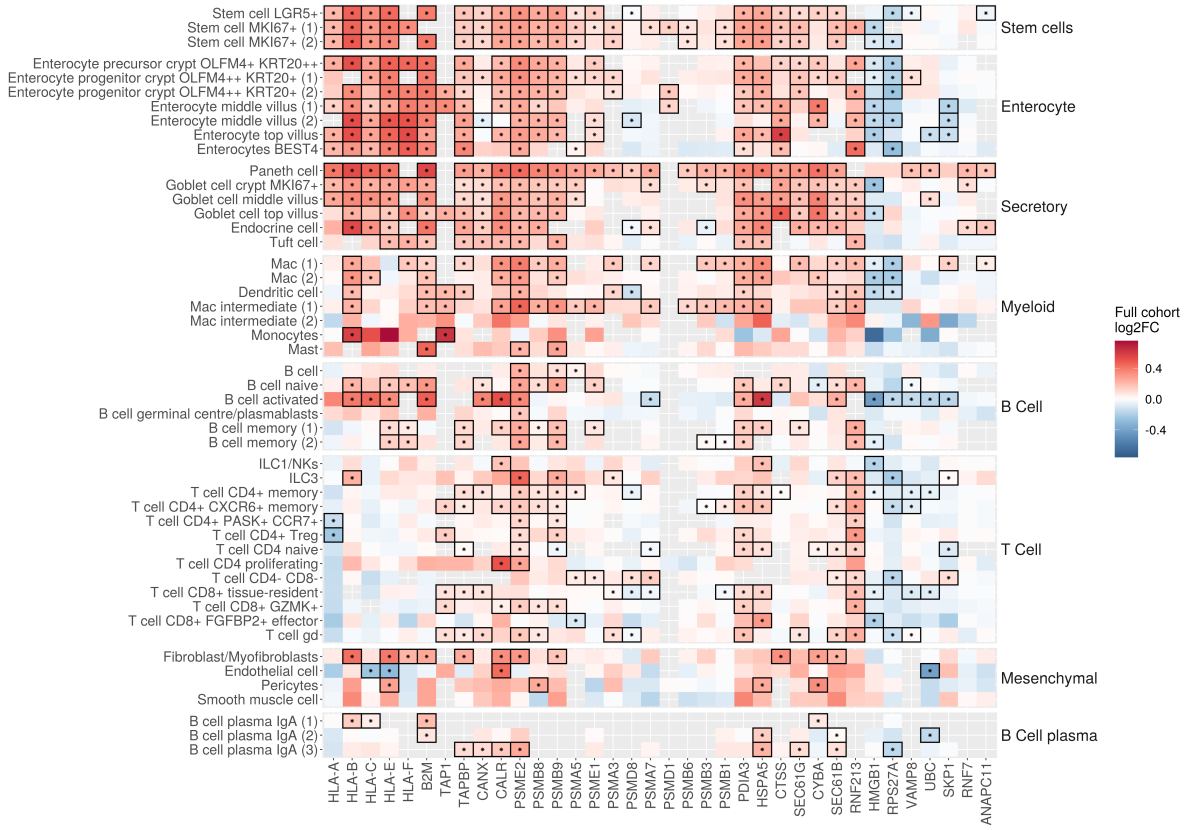
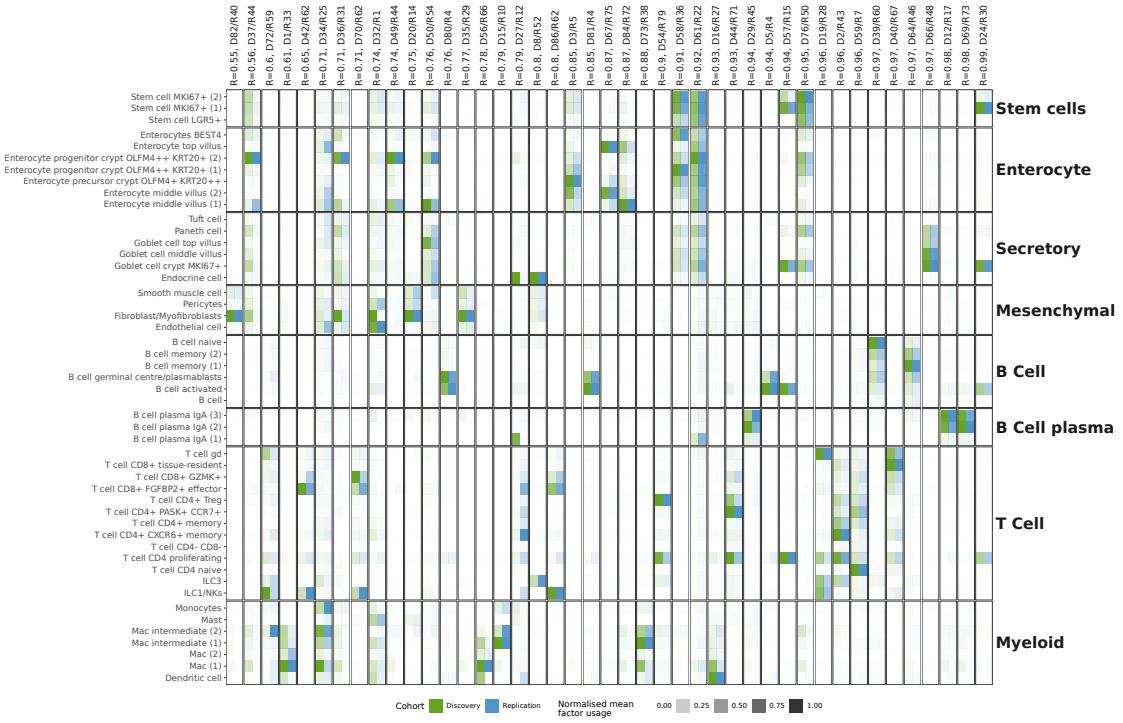


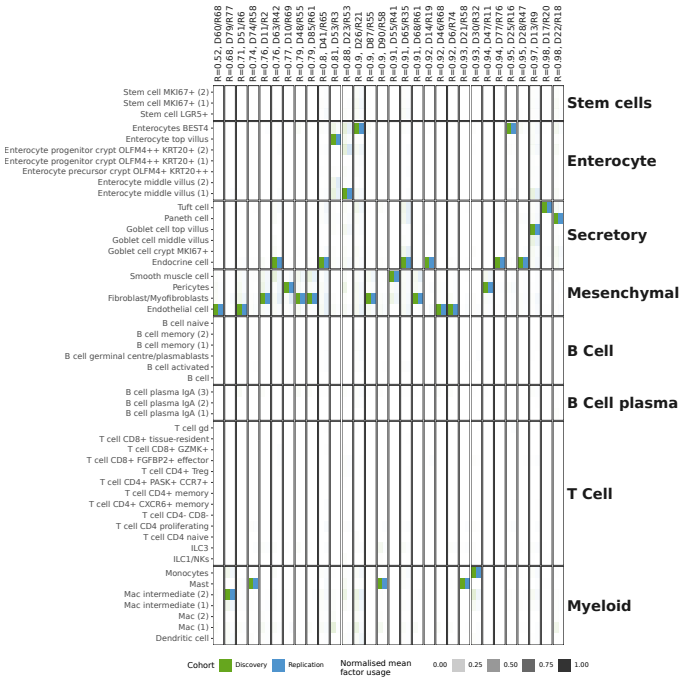
Figure 3

Figure 3. MHC class I antigen presentation and interferon signalling are widely up-regulated in epithelial cells from Crohn's disease (CD) patients with inflamed terminal ileum.
(A) Dysregulation of two immune pathways: "class I MHC mediated antigen presentation" and "interferon signalling" (dark blue text) and their respective sub-pathways (light blue text), showed enrichment (FDR $<$ 5%) across the myeloid and epithelial cell lineages. The strength of enrichment is denoted with colour. Dysregulated pathways replicable in both discovery and replication cohorts are denoted with an asterisk. (B) MHC-I mediated antigen presentation pathway genes dysregulated in full cohort (discovery and replication) with colour representing log2 fold change, and asterisk denotes significant differentially expressed genes (FDR $<$ 5%).

(A) Shared programmes of gene expression across cell types



(B) Cell type specific programmes of gene expression.



(C) Programmes associated with disease status

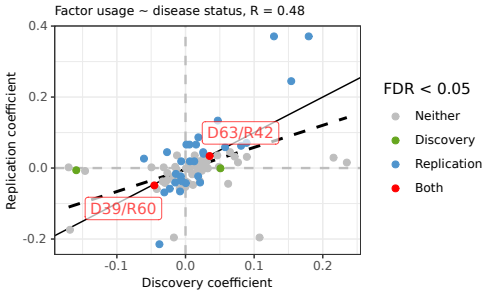
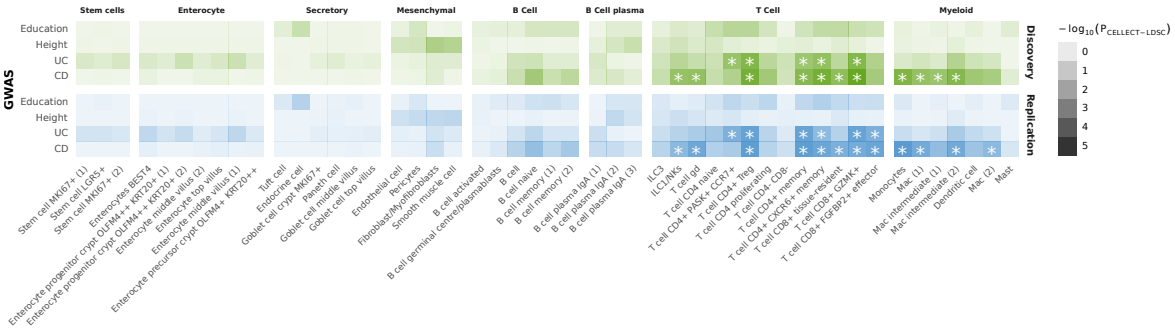


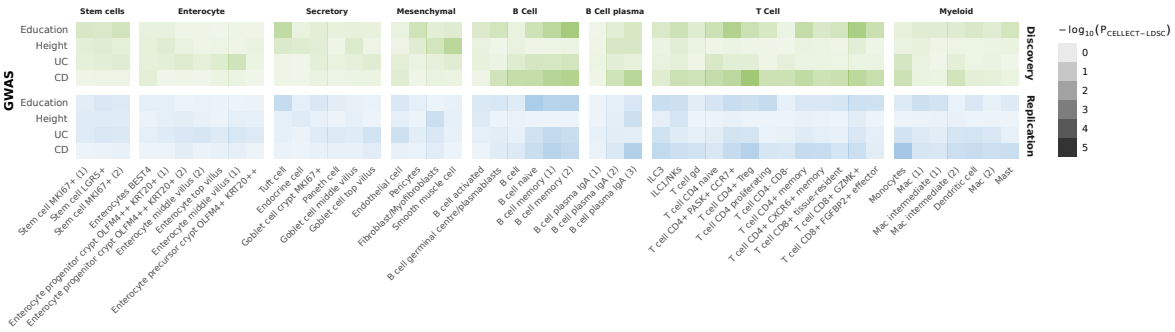
Figure 4

Figure 4. Non-negative matrix factorisation identifies programmes of gene expression.
(A) Average cell type usage of the factors shared and across cell types and (B) specific to cell types in the discovery cohort and their most correlated equivalents in the replication cohort (Pearson's $R > 0.5$), normalised to a 0-1 range. (C) Regression coefficients from fitting linear regression models between each cell factor usage and disease status. Correlated factor pairs are plotted with discovery coefficent on the x axis and replication coefficient on the y axis. Models with FDR $< 5\%$ are coloured.

(A) GWAS heritability enrichment for specifically expressed genes



(B) GWAS heritability enrichment for differentially expressed genes



(C) GWAS heritability enrichment for non-negative matrix factors

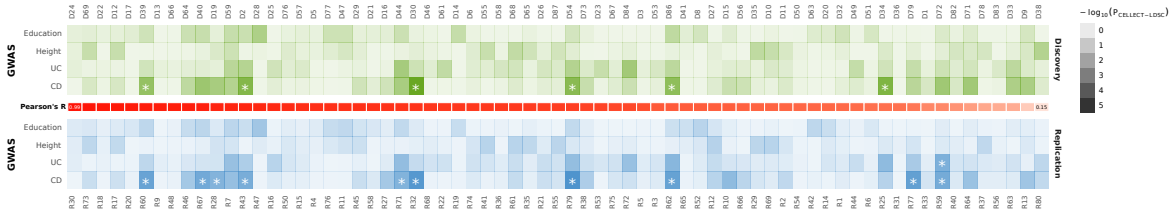


Figure 5

Figure 5. Myeloid, T cells and immune programmes are enriched for heritable risk of Crohn's disease (CD) and ulcerative colitis (UC).
Results of partitioning heritability from genome wide association study (GWAS) of CD, UC, height and educational attainment into functional categories based on (A) specifically and (B) differentially expressed genes in each cell type as well as (C) non-negative matrix factors. Categories that reach Bonferroni-corrected significance (FWER $<$ 5%) are denoted with an asterisk.

	Healthy	Crohn's disease	P-value
N	71	50	
Sex = M (%)	31 (43.7)	24 (48.0)	0.712
On Medication (%)	0 (0.0)	25 (50.0)	
Medication class (%)			
Anti-TNF	0 (0.0)	8 (16.0)	
Immunosuppressants	0 (0.0)	7 (14.0)	
Anti-TNF + Immunosuppressants	0 (0.0)	5 (10.0)	
Corticosteriods	0 (0.0)	3 (6.0)	
5-Asa Drugs	0 (0.0)	1 (2.0)	
Ustekinumab	0 (0.0)	1 (2.0)	
Age [mean (sd)]	54.93 (14.93)	39.08 (12.51)	$<$0.001

Table 1. Characterisation of participants in this study.
Healthy and Crohn's disease participants were balanced in terms of number of males and females (fisher test $p > 0.001$). Half of Crohn's disease patients were undergoing various medications including Anti-TNF and Immunosuppressants. In average, Crohn's disease patients were significantly younger (t-test $p < 0.001$) than healthy controls by 16 years.

	Discovery	Replication	P-value
N	51	70	
Crohn's disease (%)	26 (51.0)	24 (34.3)	0.092
Sex = M (%)	22 (43.1)	33 (47.1)	0.714
Age [mean (sd)]	42.12 (14.21)	52.57 (15.82)	$<$0.001

Table 2. Characterisation of discovery and replication cohorts.
Discovery cohort was balanced in number of CD and healthy participants while replication cohort contained 66% healthy patients. There was no significant difference in sex between two cohorts (fisher test $p > 0.001$).