Single-Cell RNA Sequencing of Terminal Ileal Biopsies Identifies Signatures of Crohn's Disease Pathogenesis.

Monika Krzak[1]*, Tobi Alegbe[1,2]*, D Leland Taylor[1,2]*, Gareth-Rhys Jones[3]*, Mennatallah Ghouraba[1], Michelle Strickland[1], Bradley T Harris[1], Reem Satti[1], Kenneth Arestang[4], Lucia Ramirez-Navarro[1], Nilanga Nishad[4], Kimberly Ai Xian Cheam[1], Marcus Tutert[1], Matiss Ozols[1], Guillaume Noell[1], Steven Leonard[1], Moritz J Przybilla[1], Velislava Petrova[1], Carla P Jones[1], Noor Wana[1], May Xueqi Hu[1], Jason Skelton[1], Jasmin Ostermayer[1], Yong Gu[1], Wendy Garri[4], Biljana Brezina[4], Charry Queen Caballes[4], Daniele Corridoni[2,5], Miles Parkes[4], Vivek Iyer[1], Cristina Cotobal Martin[1], Rebecca E McIntyre[1], Tim Raine[4♀], Carl A Anderson[1,2♀].

[1] Wellcome Sanger Institute, Hinxton CB10 1SA, UK
[2] Open Targets, Wellcome Genome Campus, Hinxton, CB10 1SA, UK
[3] University of Edinburgh Centre for Inflammation Research, Queens Medical Research institute, Edinburgh EH16 4TJ, UK.
[4] Department of Gastroenterology, Addenbrooke's Hospital, Cambridge University Teaching Hospitals, Cambridge CB2 0QQ, UK
[5] Sanofi R&D, The Bennet Building, Babraham Research Campus, Cambridge, CB22 3AT, United Kingdom.

* These authors contributed equally to this work

♀ These authors jointly supervised the study.

Correspondence: carl.anderson@sanger.ac.uk & tr223@cam.ac.uk

Summary

Crohn's disease (CD) is a chronic inflammatory bowel disease exhibiting substantial heterogeneity in clinical presentation and response to therapy. To explore its molecular basis, we developed IBDverse, the largest single-cell RNA sequencing (scRNA-seq) dataset of terminal ileal biopsies, profiling over 1.1 million cells from 111 CD patients and 232 healthy controls. This resource integrates discovery and replication cohorts for robust identification of CD-associated cell types, genes, and pathways. We uncovered epithelial changes marked by interferon-driven MHC-I upregulation, persisting in progenitors after macroscopic inflammation resolution. *ITGA4*$^{+}$ macrophages were identified as key inflammatory drivers, showing enriched JAK/STAT signaling and cytokine expression (IL-6, IL-12, IL-23). Heritability analysis linked inflammatory monocytes and macrophages to CD susceptibility, implicating resident and recruited immune cells in pathogenesis. These findings establish a comprehensive cellular and molecular framework for CD, offering new insights into disease mechanisms and therapeutic opportunities.

Keywords

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

Introduction

Crohn's disease (CD) is a debilitating inflammatory bowel disease (IBD) characterised by chronic relapsing and remitting inflammation of the gastrointestinal tract. The biological basis of CD is incompletely understood, although it is hypothesised to be caused by an overactive immune response to commensal gut bacteria in genetically predisposed individuals. Although inflammation is most commonly observed in the terminal ileum, CD exhibits significant heterogeneity in disease location, severity, and behaviour, both between patients and within patients over time. While therapies targeting cytokines such as TNF, IL-12 and IL-23 have improved clinical outcomes for some patients, primary non-response and secondary loss of response to treatment remains high, with 15% of CD patients requiring surgical intervention within five years of diagnosis [1], [2]. Consequently, there is an urgent need to understand better the aetiology of CD in order to broaden therapeutic opportunities.

Genetic association studies have identified over 320 loci associated with IBD susceptibility [3], [4], [5]. Together, these associations have suggested that a broad range of immune, epithelial and stromal cell genes and pathways are involved in CD pathogenesis. However, it is often challenging to draw novel biological insights from these discoveries because most of these disease-associated variants reside in the non-coding genome and it is unclear which genes and pathways they dysregulate. While functional genomic studies of disease-relevant tissues and cell types have shown utility for identifying CD effector genes, pathways and cell types [6], [7], [8], [9], [10], [3], progress has been hampered by their restriction to whole tissues or a limited number of cell types and states.

Single-cell RNA sequencing (scRNA-seq) technologies overcome this limitation by providing a high-throughput means to dissect complex tissues at the resolution of single cells and cell types. ScRNA-seq atlases of the gastrointestinal tract have already made some important discoveries, including the pathogenic remodelling of mesenchymal cells [11] and *CD8*$^+$ T cells [12] in ulcerative colitis (UC), another common form of IBD, as well as the identification of *BEST4*$^+$ enterocytes that are crucial in maintaining luminal pH in both UC and CD [13], [14], [15], and identification of metaplastic cells in IBD [16]. Falling scRNAseq costs have begun to enable cell-type resolution comparisons of gene-expression between groups of samples, for example between different gut locations [17], between disease cases and unaffected controls [14], [17] or between responders and non-responders to a drug [18]. While these early discoveries showcase the potential for scRNA-seq-based differential gene-expression studies to deliver biological insights into disease, power to detect differentially expressed genes has been constrained by the small sample sizes used to date. Furthermore, the clinical and

biological heterogeneity of IBD, along with the technical noise inherent in scRNA-seq data [19], increases the risk of false-positive associations due to confounding and further limits power to detect true associations. A large scRNA-seq study of gastrointestinal samples ascertained from hundreds of CD patients and healthy controls would thus be a key resource to identify genes, pathways and cell types that are reproducibly associated with CD and related phenotypes.

To address this need, we present IBDverse, a scRNA-seq data resource of over 1,185,000 cells isolated from terminal ileum biopsies from 111 patients with Crohn's disease and a history of current or previous terminal ileitis and 232 healthy controls. This study, which comprises both a discovery and replication cohort, is uniquely powered to identify gastrointestinal cell types, genes and pathways reproducibly associated with CD. 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 terminal ileal cell transcriptional data and associated phenotypic information from the IBDverse is available at zenodo (Data availability) and the expression atlas of gastrointestinal cell types is accessible via interactive data explorer (Data availability).

Results

Large scale single-cell sequencing of terminal ileal biopsies identifies 57 cell clusters comprising epithelial, immune and mesenchymal cell types.

Terminal ileal (TI) biopsies were collected from 343 patients during ileo-colonoscopy, comprising 232 biopsies from healthy controls and 111 biopsies from CD patients with either active inflammation in the TI or a documented history thereof (Fig. 1a). Baseline demographics of sex and smoking status were not associated with disease status ($p > 0.001$, Fisher's exact test). CD patients were significantly younger than the healthy controls by an average of 8 years (41±12 vs 49±13 years, $p<0.001$, t-test) (Table 1). The magnitude of TI inflammation was assessed using inflammatory components of the simplified endoscopic score of CD activity (SES-CD) to measure ulceration and mucosal surface involvement. The CD biopsies were subsequently classified according to endoscopic severity of disease using a score based upon three inflammatory components (inflamed surface, ulcer size, ulcerated surface) of the validated "simple endoscopic score for CD" (SES-CD), applied to the terminal ileal segment

118 by a single central reader (TI-SES-CD score). This resulted in 64 inflamed (TI-SES-CD≥3) and
119 47 uninflamed (TI-SES-CD<3) biopsies from 111 individuals (biopsies are unpaired). Clinical
120 information and metadata for the participants are provided in Table S1.
121
122 All 343 ileal biopsies were processed using a tissue dissociation protocol performed on ice
123 [20] to improve epithelial cell isolation by enhancing cell viability and minimizing cellular stress
124 compared to conventional collagenase digestion at 37°C [21]. A representative cell atlassing
125 cohort consisting of 216,376 high-quality cells (Methods) was derived from 70 randomly
126 selected samples (35 CD {uninflamed/inflamed} and 35 controls), with an average of ~4,100
127 post-QC cells from CD biopsies and ~2,000 post-QC cells from healthy biopsies (Fig. 1b).
128 Marker genes for each of the 57 unique cell clusters were identified using Wilcoxon rank-sum
129 tests, and manual annotations were performed by integrating literature references and
130 lineage-defining gene expression profiles (Fig. S1; Table S2; Methods). These 57 clusters
131 defined three major compartments, comprising eight major cell populations: epithelial cells
132 (enterocytes, secretory, and stem cells), immune cells (T, B, plasma B cells, and myeloid
133 cells), and mesenchymal cells. Positional marker genes from a spatial transcriptomic study of
134 mouse jejunum [22] were employed to annotate enterocytes and secretory cells along the
135 crypt (precursor/progenitor) and villus (bottom/middle/top) axis (Fig. S2). Gene expression
136 signatures of all 57 cell types were then used to build a cell type classifier to annotate cells
137 from the remaining 273 samples not included in the atlasing cohort.
138
139 As expected, the cellular composition was altered in CD, with overrepresentation of cell types
140 known to accumulate in intestinal inflammation. For example, *S100A8/9*$^{+}$ monocytes and
141 *CXCL9/CXCL10*$^{+}$ macrophages were effectively absent (<2%) in health (Fig. S3). In disease,
142 these cell-types comprised 25% and 7%, respectively, of all cells in the myeloid compartment,
143 in keeping with their known role in pathogen/inflammatory response [18], [23], [24], [25].
144 Similarly, four clusters of enterocytes (*GPX2*$^{+}$ progenitor crypt, *KRT20*$^{+}$ precursor crypt, *IFI27*$^{++}$
145 mid villus and *IFI27*$^{++}$ bottom villus) were all almost exclusively found in CD (Fig. S3), and
146 displayed transcriptional changes associated with barrier-response to injury. For example,
147 both the "*IFI27*$^{++}$ bottom and middle villus" clusters were denoted by high expression of
148 *CEACAM20*, a sensor of gram negative bacteria and driver of IL-8 release through SAP-1
149 phosphorylation, antimicrobial peptides (*REG1/3B*) and cytokine responsive elements (*NOS2*
150 and *HSD3B2*), though expression of these markers was consistently lower in the middle
151 versus bottom villus cluster. *HSD3B2*, a steroid synthetase essential for progesterone
152 production, has been shown to modulate local cytokine and repair mechanisms, highlighting

153 the epithelium as not simply a physical barrier, but a key component of the inflammatory
154 cascade, including extra-adrenal steroid production [26], [27].
155
156 The small sample sizes commonly used in single-cell differential gene expression studies,
157 combined with the technical and biological variability inherent in single-cell analyses of
158 complex tissues, have made it difficult to validate discoveries of differentially expressed genes
159 and pathways. This has led to claims that the field of single-cell sequencing is facing a
160 reproducibility crisis [28], [29], [30]. To directly assess the replicability of our analyses, all 343
161 samples in the study were randomly allocated to either a 'discovery' or 'replication' cohort.
162 First, the 70 samples in the atlasing cohort were equally distributed at random between the
163 two cohorts to minimize bias introduced by variation in the accuracy of cell atlasing. The
164 remaining samples not included in the atlasing cohort were then randomly assigned to either
165 cohort. Ultimately, the discovery cohort consisted of 57 CD (29 with inflamed and 28 with
166 uninflamed CD) and 114 healthy individuals, while the replication cohort included 54 CD (35
167 with inflamed and 19 with uninflamed CD) and 118 healthy controls (Fig. 2a). No significant
168 differences were observed in baseline demographics between the two cohorts (Table S3). The
169 cell-type classifier was then applied to auto-annotate all 611,992 QC passing cells in the
170 discovery cohort and 573,869 cells in the replication cohort (Methods).
171
172 Within the atlasing cohort we had both the original and auto-annotation cell labels, and were
173 able to quantify the agreement. Overall, the auto-annotation classifier performed well, with
174 similar agreement in annotation between discovery and replication cohorts, demonstrated by
175 high correlation in specifically expressed genes between the original and auto-annotated
176 clusters (Pearson's R > 0.77) (Fig. S4; Methods). As expected, the accuracy of the
177 autoannotator varied between cell-types, with 49 cell types showing an annotation
178 concordance of more than 50% (Fig. S5). Eight cell types exhibited poor annotation
179 concordance (with fewer than 50% of cells concordantly annotated) - plasmacytoid dendritic
180 cells and conventional dendritic cells type 2 (pDCs/cDC2), *GZMK*$^{+}$ and tissue-resident *CD8*$^{+}$
181 T cells, *OLFM4*$^{+}$ stem cells, *XBP1/CD38*$^{+++}$ and *XBP1/CD38*$^{++++}$ plasma B cells, and top / mid
182 villus *IFI27*$^{+}$ enterocytes. Somewhat reassuringly, misannotation was often between
183 transcriptomically similar cells. For example, misannotated plasma B *XBP1/CD38*$^{+++/+}$ cells
184 were almost exclusively classified as plasma B *XBP1/CD38*$^{++}$ cells, and misannotated *OLFM4*$^{+}$
185 stem cells were predominantly classified as *OLFM4*$^{+}$ epithelial progenitors.

186 **Identification and replication of dysregulated genes in Crohn's disease.**

187 Across each of the 57 clusters, we compared mean gene expression between inflamed CD
188 biopsies and uninflamed biopsies from healthy individuals. To directly assess replicability, this
189 was performed within both the discovery and replication cohorts independently, and across
190 the two cohorts combined (full cohort) (Fig. 2a). We found 4241 unique differentially expressed
191 genes (DEGs) in the discovery cohort, 4166 unique DEGs in the replication cohort, and 5385
192 unique DEGs in the full cohort (Table S4; false discovery rate [FDR] < 5%). As expected, the
193 total number of DEGs for each major cell population was positively correlated ($R_{Discovery}$ = 0.66,
194 $R_{Replication}$=0.52, R_{Full}=0.73) with the number of sequenced cells (Fig. S6a), albeit with variation
195 at the level of individual cell types (Fig. S6b). For example, secretory epithelial cells,
196 enterocytes and T cells had the highest number of DEGs in the full cohort (totals of 6527, 6324
197 and 4866, respectively), while plasma B cells had only 719 DEGs, despite the large number
198 of sequenced cells (~71k in full cohort).

199 We hypothesised that power to detect and replicate DEGs across our cohorts might
200 be affected by the transcriptional variability (either technical or biological) of individual cell
201 types. To assess this, we compared the gene-expression fold change estimates from our
202 differential gene expression tests between our discovery and replication cohorts (Fig. 2b).
203 When comparing the major cell populations, epithelial cells were the most replicable (mean
204 R=0.76) and plasma B cells the least (mean R=0.56). However, while eight of the top ten most
205 replicable cell types were within the epithelial cell lineage, all major cell populations showed
206 wide variation in fold-change replicability between constituent cell-types. For example, the
207 fold-changes estimates were replicable for gamma-delta T cells (R=0.82), but less so for
208 *LEF1*$^{+}$ T cells (R=0.28) or *CD8*$^{+}$ tissue-resident T cells (R=0.32) (Table S5). There was also
209 significant variation in fold-change correlation between individual cell-types in the myeloid
210 compartment, with *ITGA4*$^{+}$ macrophages demonstrating the best agreement (R=0.88) and
211 pDCs/cDC2s the least (R=0.12). Four of the ten cell types with a fold-change correlation less
212 than 0.5 (pDCs/cDC2, tissue-resident *CD8*$^{+}$ T cells, *OLFM4*$^{+}$ stem cells, *XBP1/CD38*$^{+++}$ and
213 *XBP1/CD38*$^{++++}$ plasma B cells) showed poor annotation consistency between the manual and
214 auto-annotation approaches in the atlasing cohort (fewer than 50% of cells concordantly
215 annotated). This highlights the importance of accurate and consistent cell annotation for
216 single-cell RNA-seq studies to ensure reliable differential gene expression analysis. Two
217 myeloid cell types (*SOD2*$^{+}$ monocytes and *CCL3/4*$^{+}$ *CXCL9/10*$^{+}$ macrophages) showed poor
218 fold-change correlations (R<0.5) despite high consistency between the manual and auto-
219 annotation (>98% of cells concordantly annotated). The poor replicability observed for these
220 cell-types was instead likely underpinned by imbalance in the number of cells from cases and
221 controls. Together, these factors make it difficult to accurately estimate mean gene-expression

222 for these cell-types, particularly in the controls, hindering power to detect differences in mean
223 expression correlated with disease status.
224 Ultimately, only 44% of DEGs (FDR < 5%) detected in the discovery cohort were also
225 significantly dysregulated (FDR <5%) in the same cell type, with the same direction of effect,
226 in the replication cohort. As anticipated, the correlation in gene-expression log-fold changes
227 in our discovery and replication cohort underpinned the replication rate of differentially
228 expressed genes. Consistent with this, the replicability of a given gene is highly cell-type
229 dependent, with epithelial cells, particularly precursors/progenitors and stem cells
230 demonstrating the most consistent differentially-expressed genes between cohorts.

231 **The Ileum undergoes both pan-epithelial and cell-type specific changes in CD**
232 **inflammation.**

233 To maximise power to identify biological pathways dysregulated in disease, we
234 undertook gene set enrichment analysis (GSEA) across the differential gene expression
235 results from our full cohort (Methods). To limit false-positive enrichments, we focussed this
236 analysis on the 47 cell types with greatest replicability in our differential gene expression
237 (DEG) analysis (Fig. 2b; Table S5; threshold R ≥ 0.5). Epithelial cells demonstrated the
238 strongest correlation in fold change estimates (R=0.76) of the three major cellular
239 compartments, and in the full cohort we observed both cell-type specific and pan-epithelial
240 pathway enrichments (Fig. 2c).

241 The most widespread pan-epithelial signature in CD inflamed versus healthy control
242 samples was upregulation of "IFNα/β signalling", detected in twelve epithelial cell types along
243 the entire crypt-villus axis (Fig. 2c). Leading edge analysis showed that the genes driving the
244 observed enrichments for "IFNα/β signalling", namely *IFI27, HLA-A/B/C/E/F, IFITM3, PSMB8,*
245 *STAT1* and *IRF1,* were differentially expressed across all twelve cell types (Table S6). Many
246 of these leading edge genes also featured in other significantly enriched pan-epithelial
247 pathways, including *HLA-A/B/C/E/F* and *PSMB8* in the upregulation of "Antigen processing
248 cross-presentation", "Folding-assembly, peptide loading of class I MHC", "ER phagosome"
249 and "Endosomal vacuolar pathway". To determine whether IFNα/β signalling enrichment
250 persisted in patients who had successfully resolved macroscopic inflammation, we undertook
251 differential gene expression analyses across the twelve IFNα/β-enriched cell types, comparing
252 cells from uninflamed CD biopsies to those from healthy controls (Fig. 2c). We did not observe
253 an enrichment of differentially expressed genes in the IFNα/β pathway, except for a weak but
254 statistically significant enrichment in "stem cells *MKI67*" in the uninflamed CD versus healthy
255 biopsy comparison. This suggests that the enrichment of Type I IFN signalling across epithelial

256 cell types in active CD is a widespread barrier response to injury that largely subsides after
257 repair.

258 Type I interferon (IFN-α/β) and type II interferon (IFN-γ) have long been implicated in
259 the upregulation of MHC-I expression and components of its pathway [31], [32], [33].
260 Consistent with this, genes differentially expressed between cells from CD inflamed versus
261 healthy biopsies were enriched in the "folding assembly peptide loading of class I MHC"
262 pathway, including (*CANX, CALR, TAPBP, B2M*), across 12 individual epithelial cell types
263 representing all major subtypes (Table S6). However Type II IFN responses, through the "IFNγ
264 signalling" pathway, appeared more restricted, with enrichment preferentially to enterocytes
265 along the entire crypt-villus axis, but not in other epithelial cell types, e.g. secretory cells (Fig.
266 2c). Given that multiple MHC-I pathways and associated genes were consistently enriched
267 across epithelial cell types, we speculated whether inflammation severity was associated with
268 the strength of this effect. To test this, we fit a differential expression (DE) model for CD
269 patients only stratified according to endoscopic severity of disease using the TI-SES-CD
270 score. In this analysis, "MHC-I mediated antigen processing presentation", "folding assembly
271 peptide loading MHC-I" and "antigen processing cross-presentation" pathways remained
272 enriched in all epithelial cell types except goblet cell top-villus and enteroendocrine subsets
273 (Fig. S7a). This suggests a widespread dose-dependent effect of inflammation on MHC-I
274 upregulation. Importantly, analyses comparing non-inflamed CD to those from healthy controls
275 showed that this upregulation of MHC-I antigen-presentation persisted even after the
276 resolution of inflammation (Fig. 2c).

277 Whilst MHC-I is expressed by all cells, MHC-II expression is thought to be restricted to
278 professional antigen presenting cells to limit activation of the adaptive immune system through
279 T cell receptor engagement. However, in the context of inflammation, intestinal epithelial cells
280 have been shown to upregulate MHC-II molecules [18], [13], [34], [35]. Consistent with this,
281 we also observed upregulation of some MHC-II signaling genes (*HLA-DRA, HLA-DRB1, HLA-*
282 *DPA1, HLA-DRB5, HLA-DQB1* and *HLA-DMB*) but only within specific enterocyte and stem
283 cell subsets (mid villus *ALDOB*$^{+}$, mid villus *IFI27*$^{+}$, progenitor crypt *OLMF4*$^{++}$ and *MKI67*$^{+}$ stem)
284 (Fig. S7b). Thus MHC-II signalling upregulation was less widespread across epithelial cells
285 compared to MHC-I.

286 Collectively these results suggest widespread changes across epithelial sub-types
287 relating to Type I IFN and MHC-I signalling, with these changes associated with inflammation
288 severity. Importantly, these perturbations in MHC-I signalling may persist in key progenitor
289 cells long after the initial inflammatory stimuli have been removed. Alongside these changes
290 in immune potential in epithelial cells, we also detected evidence of shifts in metabolic function

291 with evidence of alterations in pathways and genes associated with both oxidative
292 phosphorylation and glycolysis (Fig. S7c).

293 **$ITGA4^+$ macrophages are enriched for JAK/STAT signalling in ileal CD inflammation**

294 Genetic variants associated with increased expression of *ITGA4* (CD49d) in stimulated
295 monocytes have been associated with increased risk of IBD [3]. *ITGA4* is known to be
296 expressed by a range of circulating lymphocytes, including classical monocytes, where it forms
297 the α4β7 integrin that is the target of the IBD therapeutic, vedolizumab. Anti-α4β7 therapy
298 abrogates blood monocyte MADCAM-endothelium interactions, with circulating levels of $\alpha4\beta7^+$
299 classical monocytes higher, and tissue monocytes lower, in vedolizumab non-responders [36],
300 [37]. In our terminal ileum cell atlas, *ITGA4* was most highly expressed by two myeloid
301 populations, $ITGA4^+$ macrophages and $CD163^{++}$ macrophages (Fig. S1).
302 In patients with active CD, $ITGA4^+$ macrophages constituted a higher mean proportion
303 of myeloid cells within inflamed tissue (16%) compared to healthy controls (10%) (Fig. S3a).
304 These cells also demonstrated the highest reproducibility in our differential gene expression
305 analysis, with a correlation coefficient of R=0.88 (Table S5). Gene set enrichment analysis
306 revealed that differentially expressed genes in $ITGA4^+$ macrophages were enriched across a
307 wide array of cytokine pathways, including interleukins (IL) 1, 4, 6, 9, 10, 12, 13, 20, 21, 23,
308 27 and 35 (Fig. S3b). These cytokines suggest a predominant role of signalling through
309 receptors of the IL-6 and interferon (IFN) superfamilies, predominantly mediated by Janus
310 kinases (JAK) JAK1/2 and TYK2 [38], [39]. Concordantly, *JAK2* and *STAT1* were the most
311 overexpressed genes in our differential gene expression analysis, with *JAK2* and multiple
312 *STAT* isoforms (*STAT1/2/3*) identified as key mediators in the leading-edge genes across the
313 enriched pathways. Negative regulators of JAK/STAT signalling, such as *PTPN1*/*2,* were also
314 upregulated in $ITGA4^+$ macrophages and were leading edge genes in a number of enriched
315 pathways, including "IL-1 signalling", "Signalling by interleukins" and "Cytokine signalling in
316 immune system". Furthermore, we found that the expression of genes in many of these
317 enriched cytokine signalling pathways, including IL-10*,* IL-12 and IL-20, were positively
318 correlated with inflammation severity (TI-SES-CD score) in $ITGA4^+$ macrophages (Fig. S3b).
319 Building on the observed correlations between IL-10 and IL-20 signalling in $ITGA4^+$
320 macrophages and the severity of inflammation in active CD, we further investigated the
321 potential involvement of IFN super-family receptors. Notably, $ITGA4^+$ macrophages
322 upregulated genes associated with proteasome (e.g., PSMA4/5) and immunoproteasome
323 (e.g., PSMB8/9) complex formation, that are known to be dependent on IFN signalling (Fig.
324 3c). While the proteasome primarily facilitates the normal degradation of proteins, the
325 immunoproteasome is upregulated in response to pro-inflammatory signals such as Type II

326 IFN, enhancing the generation of peptides from pathogen-derived proteins for MHC-I-
327 mediated presentation to *CD8*$^+$ T cells. This extensive engagement in protein catabolism and
328 antigen presentation suggests that *ITGA4*$^+$ macrophages may play a crucial role in amplifying
329 adaptive immune responses in the local tissue environment.
330 Taken together, our findings show that *ITGA4*$^+$ macrophages are over-represented in
331 CD inflammation and preferentially express an array of cytokine signalling pathways. The
332 specific repertoire of interleukins suggest IL-6 and IFN super-family receptor signalling with
333 resultant JAK activation, particularly JAK2, with inflammation-dependent increases in *IL-10*,
334 *IL-12* and *IL-20*. These cells further demonstrate the effects of Type II IFN signalling, with
335 enrichment for many pathways involved in immunoproteasome-MHC-I communication,
336 suggesting an active role in cytokine sensing, signalling and adaptive immune-cell cross-talk.

337 **IBD-associated genes are expressed across all major cell populations of the gut.**

338

339 Genome-wide association studies have identified more than 300 regions of the human
340 genome associated with IBD susceptibility. Translating these genetic discoveries into
341 biological understanding and therapeutic hypotheses requires elucidation of the cell types in
342 which they are operative. High resolution single-cell atlases are beginning to deliver these
343 important insights, with previous studies identifying cell types that highly express some IBD-
344 associated genes [17], [40]. To build on these efforts, we compared gene-expression
345 specificity scores across the 57 cell-types in our atlas for 45 high confidence candidate effector
346 genes (Methods) from IBD genome-wide association studies (Fig. 4). All eight major cell
347 populations showed specific expression of at least one IBD effector gene, highlighting the
348 complex cellular architecture of IBD and the absence of a single, dominant cell-type or
349 pathway underlying pathology. For example, *FUT2* was specifically expressed by
350 *IFI27/KRT20*-positive enterocytes (mean specificity = 0.7) but not differentially so between CD
351 cases and controls (Table S4). *FUT2* co-ordinates synthesis of the carbohydrate HBGA,
352 responsible for ABO blood group system, but these antigens are also secreted in mucosal
353 tissue sites. Indeed, *FUT2* mutations that increase IBD risk have been associated with 'non-
354 secretor' status of these antigens, which in turn impair attachment of *Bifidobacteria* and
355 *Lactobacilli*. Abrogation of *FUT2* expression, either via transgenic approaches or naturally
356 occurring genetic polymorphisms, results in dysbiosis and increased susceptibility to
357 inflammation [41], [42]. We show that *FUT2* is predominantly expressed by enterocytes and
358 stem cells (specificity > 0.5) compared to other major cell populations (specificity < 0.36), with
359 expression greatest in precursors/progenitors and falls progressively as enterocytes mature
360 up the crypt.

361 A small subset of coding mutations in *NOD2* comprise the strongest genetic effects on
362 ileal CD in Western populations, and have been suggested to impair handling of intra-cellular
363 bacteria by myeloid cells [43], [44], [45]. We observed specific expression of *NOD2* in
364 monocytes (*S100A8/9*$^{+}$ and *SOD2*$^{+}$) and immature (e.g. *ITGA4*$^{+}$ and *CXCL9/10*$^{+}$ cells) (mean
365 specificity=0.8), but not resident (e.g. *CD163*$^{+}$ and *CD163*$^{++}$ cells) macrophages (mean
366 specificity=0.2). Across all of these cell types we found no evidence of *NOD2* differential
367 expression between CD-inflamed biopsies and those from healthy individuals (Table S4).
368 Given that these *NOD2*$^{+}$ cell types are also known to be rapidly recruited from blood as a
369 critical component of the response to intestinal damage [46], impaired bacterial handling in
370 the initial emergency response by these cells may lead to microbial persistence and represent
371 an early triggering event for chronic immune activation. Three of these *NOD2*$^{+}$ cell types
372 (*SOD2*$^{+}$ monocytes, *S100A8/9*$^{+}$ monocytes and *CXCL9/10*$^{+}$ macrophages) were also the sole
373 source of oncostatin M (*OSM*) in our data (mean specificity=0.9). Oncostatin M has been
374 suggested as a potential biomarker and therapeutic target in IBD, that induces an inflammatory
375 programme including *IL6*, *CXCL9* and *CXCL10* in stromal cells, with high pretreatment levels
376 associated with anti-TNF failure [47]. The cognate receptor, *OSMR,* has been shown to be
377 expressed on CD45^{-}EpCAM^{-}CD31^{-} cells and its expression is co-linear with *COL1A1*, *FAP*
378 and *PDPN* fibroblast markers [47], at whole tissue level. In our study, we confirm cell-specific
379 expressions of *OSMR* within fibroblast, endothelial and pericyte subpopulations of
380 mesenchymal cells (specificity=0.74, 0.72, 0.67, respectively). These observations collectively
381 demonstrate the utility of continuing to build comprehensive cellular atlases to enhance the
382 interpretation of genetic association studies of IBD.

383 **Heritability enrichment analysis identifies disease relevant cell types for CD and UC.**

384 As well as assigning putative effector cell types to specific genetic risk loci, we also
385 sought to identify cell-types of pathologic relevance based upon correlation of cell type specific
386 gene-expression scores with genome-wide maps of genetic susceptibility for CD, UC and IBD.
387 Genome-wide summary statistics for height and educational attainment were used as negative
388 controls in these heritability enrichment analyses [7], [48], [49], [50].

389 Only cell types of the innate and adaptive immune system showed enrichment of
390 heritability for CD, UC and IBD. Among innate immune cells, specifically expressed genes
391 within *S100A8/9*$^{+}$ and *SOD2*$^{+}$ monocytes, *CXCL9/CXCL10*$^{+}$ macrophages, *CD163*$^{+/++}$ and
392 cDC1s were significantly enriched for CD but not UC heritability (Fig. 5; FWER < 0.05). This
393 suggests a potential causal role for cell types involved in both inflammatory responses
394 (*S100A8/9*$^{+}$ monocytes, *CXCL9/10*$^{+}$ macrophages) and immune tolerance (*CD163*$^{+}$
395 macrophages and cDC1s) in the pathogenesis of CD. Furthermore, myeloid populations with

396 genes significantly enriched for CD heritability, including *S100A8/9*$^{+}$ monocytes, *SOD2*$^{+}$
397 monocytes and *CCL3/4*$^{+}$ *CXCL9/10*$^{+}$ macrophages, were significantly reduced in healthy
398 samples (Fig. S8; p adj < 0.05, t-test). Within the adaptive immune compartment, genes
399 specifically expressed by *CD4*$^{+}$ memory/*CXCR6*$^{+}$ T cells, Tregs, *TRGC2*$^{+}$ *CD8*$^{+}$ T cells and
400 *CD4*$^{-}$ *CD8*$^{-}$ T cells were enriched for both CD and UC heritability, whereas *CD8*$^{+}$ tissue-
401 resident T cells and gamma-delta T cells were specifically enriched in CD (Fig. 5; FWER <
402 0.05).

403 Discussion

404

405 scRNA-seq is a transformative tool for elucidating the intricate interplay of genes,
406 pathways, and cell types in chronic inflammatory diseases such as Crohn's disease. However,
407 the inherent biological complexity of these diseases results in significant inter-patient variability
408 in molecular traits like gene expression. This variability is further exacerbated by technical
409 variance—often stemming from poorly understood factors associated with biopsy collection,
410 tissue dissociation, and scRNA-seq protocols—posing substantial challenges to accurately
411 quantifying gene expression differences between groups (e.g., cases vs. controls). These
412 issues are compounded by the historically small sample sizes of scRNA-seq studies and the
413 frequent absence of replication cohorts to validate findings.
414 To overcome these limitations, we established IBDverse, the largest patient cohort to
415 date for scRNA-seq of terminal ileum biopsies, incorporating both discovery and replication
416 cohorts. We standardized biopsy collection and processing from all 343 patients at a single
417 center, employing consistent protocols for tissue dissociation, single-cell capture, sequencing,
418 and data analysis. Furthermore, we developed a custom auto-annotation model trained on our
419 dataset to reduce technical variability. Despite these rigorous measures, fewer than 50% of
420 differentially expressed genes identified in the discovery cohort were replicated in the
421 validation cohort. Replicability varied markedly across cell types, with poorly annotated or rare
422 populations exhibiting particularly low replication rates. For instance, the scarcity of *S100A8/9*$^{+}$
423 monocytes and *CXCL9/CXCL10*$^{+}$ macrophages in healthy biopsies limited our ability to
424 reliably detect differentially expressed genes. Larger studies are essential to improve detection
425 of differentially expressed genes in these disease-relevant but underrepresented cell types.
426 Additionally, the variable transcriptional plasticity of immune cells—shaped by factors
427 such as diet, inflammation severity, disease stage, and treatment regimens—likely contributes
428 to low replication rates. Notably, Crohn's disease patients in our cohort exhibited wide variation
429 in treatment histories, which likely influenced the observed transcriptomic profiles of immune
430 cells in particular. Our findings underscore the challenges of conducting scRNA-seq

differential gene-expression studies on complex, heterogeneous patient populations, and highlight the need for caution particularly when interpreting unreplicated findings from small-scale differential gene expression studies.

We identified a consistent barrier response in the intestinal epithelium, centered on Type I and Type II interferon (IFN) signalling. Genes involved in MHC-I function and Type I IFN pathways were significantly upregulated across the entire crypt-villus axis. This pan-epithelial MHC-I response to barrier damage was largely absent in other cell types and persisted even after resolution of macroscopic inflammation. These findings suggest that inflammation conditions the intestinal barrier, leaving a molecular scar within epithelial progenitors that persists after the resolution of inflammation and likely influences response to future insults. This aligns with recent work by Dennison et al. [51] in pediatric IBD, showing that loss of DNA methylation at MHC-I loci, including NLRC5, enhances MHC-I gene expression in epithelial organoids. We demonstrate that MHC-I expression remains elevated post-inflammation in adult patients, particularly in stem-like progenitor cells such as *OLFM4*$^{+}$ and *MKI67*$^{+}$ populations.

The persistent elevation of MHC-I expression following resolution of macroscopic inflammation is likely to have significant implications for barrier function, perhaps via non-canonical antigen presentation by intestinal epithelial cells (IECs) [52]. Exogenous antigens can access late endosomal compartments in IECs and colocalize with MHC-I proteins in patients with Crohn's disease (CD), raising the possibility of "cross-presentation." This process may enable IECs to present luminal antigens to CD8+ T cells, highlighting a potential role for epithelial cells in shaping immune responses. This phenomenon of cross-presentation of exogenous antigen by non-professional APCs has been reported in murine renal epithelium [53], and liver endothelium [54], 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.

Epithelial MHC-I expression driven by interferon (IFN) signalling has been observed in various tissues and disease contexts. For instance, during murine *Citrobacter* spp. infection, IFN-**γ** sensing in the intestinal epithelium facilitates both pathogen and self-antigen presentation to intraepithelial T cells via IRF1 and MHC-I, which suppresses NLRP3 macrophage inflammasome activation. Notably, mice with transgenic deletion of IFN-**γ** sensing (Ifngr^fl/flVilCre) exhibit reduced IRF1/MHC-I signalling and resistance to anti-TNF therapy [55]. Building on these findings, our data demonstrate that increased epithelial Type II IFN responses are predominantly restricted to enterocytes along the crypt-villus axis and resolve following tissue repair.

467 Macrophages are among the most abundant leukocytes in the gastrointestinal tract,
468 forming a dense network around the epithelial barrier and playing critical roles in intestinal
469 health [46]. Their dysregulation is strongly associated with chronic inflammatory diseases such
470 as IBD [56], [46]. During active disease, intestinal monocytes and their macrophage progeny
471 accumulate in large numbers, producing pro-inflammatory cytokines like TNF, IL-1β, IL-6, and
472 IL-23 [57], [58], [59]. Although macrophages are hypothesized to be key targets of IBD
473 therapies [46], their heterogeneity, especially in disease states, remains poorly understood. In
474 this study, we identified a transcriptionally distinct population of *ITGA4*$^{+}$ macrophages that
475 become more abundant during intestinal inflammation. Differentially expressed genes in these
476 cells are enriched in nine cytokine pathways, including IL-6, IL-12, and IL-23. This response
477 is underpinned by a core JAK/STAT gene signature, with JAK2 and STAT1/2/3 being
478 significantly differentially expressed. Given that JAK inhibitors, which target multiple cytokine-
479 dependent pathways (e.g., IL-6, IL-10, and IL-23), are now effective therapies for IBD,
480 understanding their cellular targets is critical. Importantly, macrophages from IBD patients
481 carrying disease associated polymorphisms in JAK2 display enhanced JAK2 expression and
482 NOD2-induced JAK2 phosphorylation, and amplified cytokine signalling [60]. Moreover, JAK2-
483 deficient macrophages fail to upregulate MHC-I proteins in response to Type I IFN stimulation,
484 increasing susceptibility to infection and inflammation [61], [62]. We identify *ITGA4*$^{+}$
485 macrophages as a novel, preferential source of JAK/STAT signalling and downstream
486 cytokine responses in Crohn's disease, highlighting their potential as therapeutic targets.
487 Specifically expressed genes from multiple cell-types within the innate and adaptive
488 immune system were enriched with heritability, highlighting their involvement in disease
489 pathogenesis. The strong enrichment of CD heritability among *CXCL9/10*$^{+}$ macrophages,
490 *SOD2*$^{+}$ and *S100A8/9*$^{+}$ monocytes is particularly noteworthy given both populations are
491 significantly expanded in disease and highly express inflammatory cytokines and known CD
492 susceptibility genes such as *NOD2, PTAFR* and *LRRK2*. Unfortunately, we were unable to
493 detect many reproducible differentially expressed genes in these important cell types, likely
494 due to the rarity of these cells in biopsies from healthy individuals. Whilst we did not identify
495 heritability enrichment within non-immune cell populations, this does not mean that these cells
496 do not contribute to disease pathology. We note that heritability enrichment in non-immune
497 cells has been reported previously, for example within M cells in UC [63]. Our data shows that
498 several likely effector genes within IBD associated loci are specifically expressed by non-
499 immune cells, including *RNF186 [64]*, *FUT2* [65], *PDLIM5 [4]* and *HNF4A* [66]. The
500 underrepresentation of stromal cells in our atlas due to a combination of using pinch biopsies
501 and digestive enzymes that favour epithelial cell capture likely also reduces our power to
502 detect enrichments in these cell types.
503

In summary, we present IBDverse, the largest single-cell RNA sequencing (scRNA-seq) dataset of terminal ileal biopsies, comprising over 1.1 million cells from 111 Crohn's disease (CD) patients and 232 healthy controls. Using both discovery and replication cohorts, the study identifies and validates differentially expressed genes, pathways, and cell types associated with CD. We find widespread epithelial changes driven by interferon signalling and persistent MHC-I upregulation post-inflammation, and a distinct population *ITGA4* expressing macrophages that are major contributors to JAK/STAT signalling and cytokine production. Heritability analysis highlights the involvement of immune cell populations, such as *CXCL9/10*$^{+}$ macrophages and *S100A8/9*$^{+}$ monocytes, in CD pathogenesis. These results provide a valuable resource for understanding CD mechanisms and identifying potential therapeutic targets. An open-access portal for navigating our single-cell data resource will be launched soon (Data availability).

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. We also thank Kylie R James for her assistance with providing cell type markers.

Author contributions

Statistical analysis and manuscript drafting, M.K., T.A., D.L.T. and G.R.J.;
Sample collection, K.A., W.G., B.B., and C.Q.C.; Sample processing, M.H.G., M.S., N.W., J.S., J.O., M.X.H., K.A.C., and R.E.M.; Clinical information, N.N. ; Critical discussion M.J.P., L.R.N, R.E.M., R.S., V.P., C.P.J., C.C., and M.P.; Data processing M.T., M.O., G.N., S.L., V.I., and Y.G; Writing – Review & Editing, D.C., B.T.H., R.E.M., T.R. and C.A.A.; Conceptualization and Supervision, T.R., C.A.A.

536 ## Declaration of interests

537 C.A.A. has received research grants or consultancy/speaker fees from Genomics plc,
538 BridgeBio, GSK and AstraZeneca. T.R. has received research/educational grants and/or
539 speaker/consultation fees from Abbvie, Arena, Aslan, AstraZeneca, Boehringer-Ingelheim,
540 BMS, Celgene, Ferring, Galapagos, Gilead, GSK, Heptares, LabGenius, Janssen, Mylan,
541 MSD, Novartis, Pfizer, Sandoz, Takeda and UCB. D.C. is now an employee at AstraZeneca
542 and R.M. is an employee at Relation Therapeutics.
543

544 ## Materials and Methods

545 ### Sample ascertainment

546 This study was approved by the National Health Service (NHS) Research Ethics Committee
547 (Cambridge South, REC ID 17/EE/0338). Written informed consent was given by all
548 participants.
549
550 Individuals undergoing routine endoscopic assessment were recruited at Addenbrooke's
551 hospital, Cambridge, UK. Clinical information and metadata for the participants are provided
552 in Table S1. All CD participants classified as 'inflamed' had a confirmed history of CD and
553 macroscopic evidence of terminal ileal inflammation from tissue sampled during the biopsy.
554 All control participants were undergoing endoscopic assessment or surveillance for healthy
555 and non-cancer related reasons (e.g., history of iron deficiency anaemia, family history of
556 colorectal cancer). Control participants did not have macroscopic evidence of intestinal
557 inflammation, a personal history of cancer, and were not in receipt of corticosteroids or any
558 other immune modulating therapy. Patients who were taking probiotics or antibiotics were
559 excluded. Patients of non-European ancestries were also excluded to reduce confounding.
560 Pinch-biopsies of the terminal ileum were collected from all participants and deposited into
561 pre-chilled Hanks Balanced Salt Solution (HBSS) without Mg^{2+}, Ca^{2+}, or phenol red. Samples
562 were placed on ice and immediately transferred to the Sanger Institute.

563 ### Single-cell RNA isolation and sequencing

564 Terminal ileal biopsies were dissociated using a single-step digestion protocol on ice to
565 release all major intestinal cell types present in the biopsy (epithelial, immune, and stromal)
566 without stressing the cells. First, the biopsies were mechanically minced and pipetted to
567 release immune cells (fraction 1) from the lamina propria and the remaining tissue chunks
568 were transferred to $HBSS^{-/-}$ containing 2 mM EDTA, 0.26 U/µl serine endoprotease isolated

569 from *Bacillus licheniformus* (Sigma, P5380), 5 μM QVD-OPh (Abcam, ab141421), and 50 μM
570 Y-27632 dihydrochloride (Abcam, ab120129). Tissue chunks were pipetted regularly during a
571 30 minute incubation on ice to release epithelial and stromal cells (fraction 2). The cells from
572 both fractions are washed, centrifuged, and then incubated for 10 minutes at room
573 temperature in Hank's Balanced Salt Solution (HBSS) with Mg^{2+}, Ca^{2+}, and without phenol
574 red—including 5 mM $CaCl_2$, 1.5U/μl collagenase IV (Worthington, LS004188), and 0.1 mg/ml
575 DNase I (Stem Cell Technologies, 07900). The cells were then filtered (30 μm; CellTrics 04-
576 0042-2316), washed, and centrifuged before being incubated for 3 minutes at room
577 temperature in the red blood cell lysis buffer (ACK lysis buffer; Gibco, A10492). Two final
578 washes and centrifugations were performed before a final filtration (40 μm) and manual cell
579 counting (haemocytometer, NanoEnTek, DHC-N01).
580
581 Single-cell RNA sequencing was undertaken using 3' 10X Genomics kits (v3.0 and v3.1)
582 according to the manufacturer's instructions. All samples sequenced under kit version v3.1
583 had dual indexes, samples sequenced under kit version v3.0 had either single or dual indexes.
584 We targeted 6,000 cells for CD participants and 3,000 cells for controls to account for the
585 increased cellular heterogeneity in CD biopsies. Since the proportions of immune cells vary
586 with inflammation status, we altered the ratio of cells from fractions 1 and 2 in an attempt to
587 make the representation of cell types more equal. Viability of the mixed populations was
588 92±9% (mean ± S.D.) according to Trypan blue staining. Libraries were sequenced using a
589 HiSeq4000 sequencer (Illumina; N_{CD}=20, $N_{control}$=4) or NovaSeq S4 XP sequencer (Illumina;
590 N_{CD}=91, $N_{control}$=228) with 100bp paired-end reads, targeting 50,000 reads per cell. We
591 compared the fraction of reads mapped confidently to the transcriptome (output metric from
592 CellRanger) within CD and healthy participants and found no difference between sequencers
593 (minimum p-value > 0.05, Wilcoxon rank sum test).

594 Single-cell RNA-seq processing and quality control procedures

595 CellRanger v3.0.2 was used to demultiplex reads, align reads to GRCh38 with Ensembl
596 version 93 transcript definitions (GRCh38-3.0.0 reference file distributed by 10X Genomics),
597 and generate cell by gene count matrices. CellBender v2.1 [67] was then applied to identify
598 droplets containing cells and adjust the raw counts matrix for background ambient transcript
599 contamination. For training, CellBender requires a rough estimate of the number of droplets
600 containing cells (cell droplets) and the number of droplets without cells (empty droplets)
601 derived from the UMI curve—the rank ordering droplet barcodes according to total UMI counts
602 (x axis) by the total number of UMI counts per droplet (y axis). The UMI curve was calculated
603 from droplets with a UMI count >1,000, and the threshold estimated using the "barcoderanks-

604 inflection" procedure from DropletUtils v1.9.16 [68]. To estimate the number of empty droplets,
605 we calculated the UMI curve as described above, selected droplets with a UMI count between
606 250 and 10, and estimated the threshold by performing both the "barcoderanks-inflection" and
607 "barcoderanks-knee" procedure from DropletUtils—using 1/3rd of the distance between the
608 two estimates as the final threshold. CellBender was run with default parameters except for
609 excluding droplets with <10 UMI counts (--low-count-threshold) and using 300 epochs with a
610 learning rate of $1x10^{-7}$. The final counts matrix was adjusted for the ambient transcript
611 signature at a false positive rate of 0.1. Next, multiplets were identified and removed using
612 scrublet v0.2.1 [69], simulating 100,000 multiplets and calculating the multiplet threshold using
613 the threshold_li function from the scikit-image package v0.17.2 [70], initialised using the
614 threshold_otsu function. The reported sex of each sample was verified by generating
615 pseudobulk expression matrices and comparing the expression of *XIST* to the mean
616 expression of all genes on the Y chromosome.

617 *De novo* cell type identification

618 The atlasing cohort was used to identify cell types and fit a model to automatically predict cell
619 types across the entire dataset. First, additional filters were applied to ensure only the highest
620 quality cells were used for *de novo* clustering. Cells with fewer than 100 genes expressed at
621 ≥ 1 count, or where the percentage of counts originating from the mitochondrial genome
622 (https://www.genenames.org/data/genegroup/#!/group/1972) was > 50, were removed. Next,
623 an isolation forest (scikit-learn v0.23.2) was used to remove outlier cells based on (i) the
624 percentage of counts originating from the mitochondrial genome, (ii) the total number of UMI
625 counts per cell, (iii) the number of genes expressed (>=1 count) per cell. These metrics were
626 selected following the recommendations in [71].
627
628 Subsequent processing and management of the expression data was performed using scanpy
629 v1.6.0 [72]. Genes expressed (>=1 count) in five or fewer cells across the whole dataset were
630 removed (sc.pp.filter_genes with min_cells=5). To account for variable sequencing depth
631 across cells, unique molecular identifier (UMI) counts were normalised by the total number of
632 counts per cell, scaled to counts per 10,000 (CP10K; sc.pp.normalise_per_cell), and the
633 CP10K expression matrix (ln[CP10K+1]; sc.pp.log1p) was log-transformed.
634
635 To perform dimensionality reduction, the 2,000 most variable genes across samples were
636 selected by (i) calculating the most variable genes per sample and (ii) selecting the 2,000
637 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) 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. To select the number of PCs for subsequent analyses, we used a scree plot [73] and calculated the “knee/elbow” derived from the variance explained by each PC using the kneedle estimator v0.7.0 [74]. 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 [75] 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 [76] 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 [77]. The final cluster classifications were chosen to achieve a minimum MCC of > 0.75 across all clusters, with a resolution of 3.25 selected to meet this criterion (Fig. S9). At this resolution, all clusters met the threshold except for cluster 40, which exhibited MCC < 0.75 at many resolutions so was excluded (Fig. S9). This adjustment yielded a total of 57 clusters.

Cell type annotation

To determine the cell type identity of the 57 clusters, marker genes for each cluster were identified using the Wilcoxon rank-sum test (sc.tl.rank_genes_groups with method='wilcoxon') to compare the gene expression of each cell type to all other cell types and rank genes according to differences in expression. Highly discriminative marker genes with a Bonferroni-corrected p-value <0.05 were then used 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. 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.

673 Within epithelial cells, we identified three distinct stem cell populations: *OLFM4*$^{+}$ stem cells,
674 *OLFM4*$^{+}$ *LGR5*$^{+}$ stem cells, and proliferating *MKI67*$^{+}$ stem cells. Among enterocytes, we
675 identified progenitor cells marked by *OLFM4* and *GPX2*, precursor cells expressing *KRT20*,
676 and a range of enterocytes expressing *IFI27*. We further distinguished enterocytes along the
677 crypt-villus axis (crypt, middle, top) based on signature genes such as *ALPI, APOA4,* and
678 *APOC3* [22]. In the secretory cell lineage, we identified goblet cells (*CLCA1, FCGBP, MUC2*),
679 including proliferating *MKI67*$^{+}$ and *BCAS1*$^{+}$ goblet cells, as well as goblet cells positioned along
680 the crypt-villus axis, marked by *EGFR, KLF4, NT5E*, and *SLC17A5*. Additional cell types
681 included enteroendocrine cells (*NTS, PYY, GCG*), enterochromaffin cells (*TPH1, CES1*), and
682 tuft cells (*PLCG2, PTGS1, LRMP*). A summary of these markers and their expression across
683 cell types is visualised in Fig. S1.
684
685 Within immune cells we identified monocytes expressing *S100A8/9, SOD2, and CXCL9/10,*
686 macrophages with positive expression of ITGA4, resident macrophage populations (*CD163,*
687 *MAF, C1QA/B/C*), conventional dendritic cells type 1 cDC1 (*XCR1, BATF3*), a mix of
688 plasmacytoid dendritic cells and conventional dendritic cells type 2 (pDC/cDC2) determined
689 by (*IRF4, ZEB1, FLT3*) and mast cells (*MS4A2, TPSAB1*) (Fig. S1). Additionally, we identified
690 thirteen distinct T cell populations, including *CD4*$^{+}$ and *CD8*$^{+}$ T cells, innate lymphoid cells
691 (ILCs, marked by *IL1R1, ALDOC, LSTI*), and gamma-delta T cells (*TRGC1, TRDC, GZMA/B*).
692 Within the *CD4*$^{+}$ T cell subset, we characterized naive T cells (*SELL, CCR7*), *LEF1*$^{+}$ and
693 *PASK*$^{+}$ expressing *CD4*$^{+}$ T cells, regulatory T cells (Tregs, marked by *FOXP3, TIGIT*), and two
694 populations of double-negative (*CD4*$^{-}$ *CD8*$^{-}$) T cells, one of which shows elevated *MBLN1*
695 expression. In the *CD8*$^{+}$ T cell subset, we identified two populations of tissue-resident *CD8*$^{+}$ T
696 cells expressing *TRGC2,* as well as a distinct population of GZMK^{+} expressing *CD8*$^{+}$ T cells
697 (Fig. S1).

698 Among B cells, we identified *FAU* expressing B cells, activated B cells (*CKS1B, STMN1*),
699 naive B cells (*IGHD/M, FCER2*), and germinal center/plasmablast B cells (*CD19, CD38,*
700 *TCL1A*). Additionally, we observed a gradient of plasma cells with varying levels of *XBP1* and
701 *CD38* expression (Fig. S1).

702 Lastly, we identified three mesenchymal cell populations: fibroblasts (*COL1A1/2, COL3A1*),
703 endothelial cells (*PECAM1, VWF*) expressing *ACKR1*, and pericytes (*PDGFRB, CSPG4*) (Fig.
704 S1).

705 We did not detect a distinct group of neutrophil and eosinophil cells - as previously well
706 documented, they are poorly represented due to the limited ability of the 10X scRNA-seq
707 process to capture granulocytes [78].

Crypt-villus score

We positioned epithelial cells along the crypt-villus axis, identifying top-villus epithelial cells as well as enterocyte precursors/progenitors and goblet cells located at the crypt base. Stem cells and enterocytes were scored based on expression of genes such as *APOA4, APOC3, ALPI, PKIB, PMP22,* and *SLC28A2*, derived from spatial transcriptomics data that characterize crypt and villus intestinal cells [22]. Similarly, we applied signature genes *EGFR, KLF4, NT5E,* and *SLC17A5* to score secretory cells across the crypt-villus axis.

Automatic cell type annotation

To annotate cell types across all samples, we used Celltypist v1.6.2 [79] to train a classification model based on our identified clusters. This trained Celltypist model was then applied to assign cell type labels to all 343 scRNA-seq samples, following the initial processing steps detailed in the "Single-cell RNA-seq processing and quality control" section. Cells with a Celltypist confidence score below 0.5 were excluded from subsequent analyses.

Differential gene expression analysis

For each cell type, we tested for association of gene expression with CD disease status or inflammation severity (TI-SES-CD score) 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) [80]. 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 $\sigma^2_{\gamma k}$. In all DE comparisons, whether contrasting CD inflamed or uninflamed with healthy controls or examining association of gene expression to TI-SES-CD score - sex, age (binned into groups of five), cell mitochondrial percentage (technical covariate associated with cellular stress), and cell complexity i.e., the number of genes detected per cell [80], [14]

741 were included as fixed effect variables and individual as a random effect to control for
742 pseudoreplication bias [81]. The Benjamini-Hochberg procedure [82] was used to control for
743 multiple testing across all cell types, and p-values were obtained from the hurdle model,
744 derived from the summed χ^2 null distributions of the discrete (Z_i) and continuous (Y_i)
745 components, as described in [80]. To increase the speed of each test, genes with an average
746 CP10K of <1 in that cell type were removed prior to fitting models for each cell type.

747 Gene set enrichment analysis

748 Gene set enrichment analyses were performed using GSEA v1.17.1 [83] with default
749 parameters to identify pathways enriched among differentially expressed genes. Pathways
750 were obtained from the reactome v76 gene pathway database [84] as part of the molecular
751 signatures database (MSigDB) v7.4 [85]. Z-scores from the CD vs control differential gene
752 expression hurdle model, were used as input for enrichment analyses.
753

754 Prioritisation of IBD effector genes

755 Forty four genes likely to be perturbed in IBD were identified from within IBD-associated loci
756 based on a several criteria, including but not limited to 1) presence of a coding mutation fine-
757 mapped down to single variant resolution, 2) detailed and convincing functional follow-up work
758 that established the causality of the gene or 3) the protein encoded by the gene plays a major
759 role in a pathway that is targeted by an existing IBD therapy. Note, it is typically not
760 straightforward to identify disease effector genes from within GWAS loci and this challenge
761 remains a major focus for the field of complex disease genetics. While our list of 45 likely IBD
762 effector genes is undoubtedly greatly enriched for true IBD effector genes, false-positives
763 could still remain.

764 Specifically expressed genes

765 We identified specifically expressed genes (SEGs) for each cell type using CELLEX v1.2.1
766 package [48]. CELLEX calculates specifically expressed gene scores using four
767 complementary approaches that include Gene Enrichment Score [86], Expression Proportion
768 [87], Normalized Specificity Index [88] and Differential Expression T-statistic - the package
769 produces a normalised mean of these four metrics which we used as our Specificity score.

770 Heritability analysis

771 Heritability enrichment analysis was performed using the CELLECT v1.3.0 workflow [48]. This
772 workflow deploys stratified LD score regression (S-LDSC) [7] to identify cell types with features
773 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 [3] and as negative controls, we used genetic summary statistics for height [49] and [50]. For the gene scores, we used cell type specific gene scores with Bonferroni correction to control for the total number of tests across all cell types.

Data availability

The datasets generated and analyzed during this study will be made publicly available upon publication of the manuscript in a peer-reviewed journal.

Code availability

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

Supplementary Figures and Tables

Fig. S1. Marker gene expression used to curate annotations within the terminal ileum atlas.

Fig. S2. Epithelial cell types represent the crypt-villus axis differentiation.

Fig. S3. Cell-type proportions across healthy and CD samples in the atlasing cohort.

Fig. S4. Concordance of gene specificities across discovery and replication datasets.

Fig. S5. Accuracy in re-annotating the atlas cohort.

Fig. S6. Differentially expressed genes between CD inflamed and healthy samples across all 57 cell types.

Fig. S7. Dysregulated pathways in CD versus healthy epithelial cells.

Fig. S8. Myeloid cell types enriched for CD heritability are found predominantly in CD gut biopsies.

Fig. S9. Optimisation of cluster resolution for cell-type identification.

Table S1. Clinical information for the IBDverse samples.

Table S2. Manually curated list of marker genes for cell types in the atlasing cohort dataset.

Table S3. Demographics of healthy and disease samples across cohorts.

Table S4. Significantly differentially expressed genes within each of the identified 57 cell types in the auto-annotated discovery, replication and full cohort datasets.

Table S5. Correlation between gene-expression fold changes from differential gene expression tests between our discovery and replication cohorts.

Table S6. Dysregulated pathways in CD inflamed epithelial cells.

837 Bibliography

838 1 Jenkinson, P.W. *et al.* (2020) Temporal Trends in Surgical Resection Rates and Biologic
839 Prescribing in Crohn's Disease: A Population-based Cohort Study. *J Crohns Colitis* 14,
840 1241–1247
841 2 Tsai, L. *et al.* (2021) Contemporary Risk of Surgery in Patients With Ulcerative Colitis
842 and Crohn's Disease: A Meta-Analysis of Population-Based Cohorts. *Clin.*
843 *Gastroenterol. Hepatol.* 19, 2031-2045.e11
844 3 de Lange, K.M. *et al.* (2017) Genome-wide association study implicates immune
845 activation of multiple integrin genes in inflammatory bowel disease. *Nat. Genet.* 49,
846 256–261
847 4 Sazonovs, A. *et al.* (2022) Large-scale sequencing identifies multiple genes and rare
848 variants associated with Crohn's disease susceptibility. *Nat. Genet.* 54, 1275–1283
849 5 Liu, Z. *et al.* (2023) Genetic architecture of the inflammatory bowel diseases across East
850 Asian and European ancestries. *Nat. Genet.* 55, 796–806
851 6 Soskic, B. *et al.* (2019) Chromatin activity at GWAS loci identifies T cell states driving
852 complex immune diseases. *Nat. Genet.* 51, 1486–1493
853 7 Finucane, H.K. *et al.* (2018) Heritability enrichment of specifically expressed genes
854 identifies disease-relevant tissues and cell types. *Nat. Genet.* 50, 621–629
855 8 Bossini-Castillo, L. *et al.* (2022) Immune disease variants modulate gene expression in
856 regulatory CD4+ T cells. *Cell Genomics* 2, None
857 9 Nasrallah, R. *et al.* (2020) A distal enhancer at risk locus 11q13.5 promotes suppression
858 of colitis by Treg cells. *Nature* 583, 447–452
859 10 Fairfax, B.P. *et al.* (2014) Innate immune activity conditions the effect of regulatory
860 variants upon monocyte gene expression. *Science* 343, 1246949
861 11 Kinchen, J. *et al.* (2018) Structural remodeling of the human colonic mesenchyme in
862 inflammatory bowel disease. *Cell* 175, 372-386.e17
863 12 Corridoni, D. *et al.* (2020) Single-cell atlas of colonic CD8+ T cells in ulcerative colitis.
864 *Nat. Med.* 26, 1480–1490
865 13 Parikh, K. *et al.* (2019) Colonic epithelial cell diversity in health and inflammatory bowel
866 disease. *Nature* 567, 49–55
867 14 Smillie, C.S. *et al.* (2019) Intra- and Inter-cellular Rewiring of the Human Colon during
868 Ulcerative Colitis. *Cell* 178, 714-730.e22
869 15 Elmentaite, R. *et al.* (2020) Single-Cell Sequencing of Developing Human Gut Reveals
870 Transcriptional Links to Childhood Crohn's Disease. *Dev. Cell* 55, 771-783.e5
871 16 Oliver, A.J. *et al.* (2024) Single-cell integration reveals metaplasia in inflammatory gut
872 diseases. *Nature* 635, 699–707
873 17 Kong, L. *et al.* (2023) The landscape of immune dysregulation in Crohn's disease
874 revealed through single-cell transcriptomic profiling in the ileum and colon. *Immunity* 56,
875 444-458.e5
876 18 Martin, J.C. *et al.* (2019) Single-Cell Analysis of Crohn's Disease Lesions Identifies a
877 Pathogenic Cellular Module Associated with Resistance to Anti-TNF Therapy. *Cell* 178,
878 1493-1508.e20
879 19 Jia, C. *et al.* (2017) Accounting for technical noise in differential expression analysis of
880 single-cell RNA sequencing data. *Nucleic Acids Res.* 45, 10978–10988
881 20 Adam, M. *et al.* (2017) Psychrophilic proteases dramatically reduce single-cell RNA-seq
882 artifacts: a molecular atlas of kidney development. *Development* 144, 3625–3632
883 21 Uniken Venema, W.T.C. *et al.* (2022) Gut mucosa dissociation protocols influence cell

884 type proportions and single-cell gene expression levels. *Sci. Rep.* 12, 9897
885 22 Moor, A.E. *et al.* (2018) Spatial Reconstruction of Single Enterocytes Uncovers Broad
886 Zonation along the Intestinal Villus Axis. *Cell* 175, 1156-1167.e15
887 23 Garrido-Trigo, A. *et al.* (2023) Macrophage and neutrophil heterogeneity at single-cell
888 spatial resolution in human inflammatory bowel disease. *Nat. Commun.* 14, 4506
889 24 Thomas, T. *et al.* (2024) A longitudinal single-cell atlas of anti-tumour necrosis factor
890 treatment in inflammatory bowel disease. *Nat. Immunol.* DOI: 10.1038/s41590-024-
891 01994-8
892 25 Stankey, C.T. *et al.* (2024) A disease-associated gene desert directs macrophage
893 inflammation through ETS2. *Nature* 630, 447–456
894 26 Stegk, J.P. *et al.* (2009) Expression profiles of human 11beta-hydroxysteroid
895 dehydrogenases type 1 and type 2 in inflammatory bowel diseases. *Mol. Cell.*
896 *Endocrinol.* 301, 104–108
897 27 Noti, M. *et al.* (2010) TNF suppresses acute intestinal inflammation by inducing local
898 glucocorticoid synthesis. *J. Exp. Med.* 207, 1057–1066
899 28 Gibson, G. (2022) Perspectives on rigor and reproducibility in single cell genomics.
900 *PLoS Genet.* 18, e1010210
901 29 Squair, J.W. *et al.* (2021) Confronting false discoveries in single-cell differential
902 expression. *Nat. Commun.* 12, 5692
903 30 Lähnemann, D. *et al.* (2020) Eleven grand challenges in single-cell data science.
904 *Genome Biol.* 21, 31
905 31 Gobin, S.J. *et al.* (1999) Transactivation of classical and nonclassical HLA class I genes
906 through the IFN-stimulated response element. *J. Immunol.* 163, 1428–1434
907 32 Abarca-Heidemann, K. *et al.* (2002) Regulation of the expression of mouse TAP-
908 associated glycoprotein (tapasin) by cytokines. *Immunol. Lett.* 83, 197–207
909 33 Johnson, D.R. and Pober, J.S. (1990) Tumor necrosis factor and immune interferon
910 synergistically increase transcription of HLA class I heavy- and light-chain genes in
911 vascular endothelium. *Proc Natl Acad Sci USA* 87, 5183–5187
912 34 Maddipatla, S.C. *et al.* (2023) Assessing Cellular and Transcriptional Diversity of Ileal
913 Mucosa Among Treatment-Naïve and Treated Crohn's Disease. *Inflamm. Bowel Dis.*
914 29, 274–285
915 35 Kanke, M. *et al.* (2022) Single-Cell Analysis Reveals Unexpected Cellular Changes and
916 Transposon Expression Signatures in the Colonic Epithelium of Treatment-Naïve Adult
917 Crohn's Disease Patients. *Cell. Mol. Gastroenterol. Hepatol.* 13, 1717–1740
918 36 Li Yim, A.Y.F. *et al.* (2023) Single-cell characterization of peripheral blood mononuclear
919 cells from Crohn's disease patients on vedolizumab. *medRxiv* DOI:
920 10.1101/2023.06.23.23291732
921 37 Mennillo, E. *et al.* (2024) Single-cell and spatial multi-omics highlight effects of anti-
922 integrin therapy across cellular compartments in ulcerative colitis. *Nat. Commun.* 15,
923 1493
924 38 Spiewak, T.A. and Patel, A. (2022) User's guide to JAK inhibitors in inflammatory bowel
925 disease. *Current Research in Pharmacology and Drug Discovery* 3, 100096
926 39 Coskun, M. *et al.* (2013) Involvement of JAK/STAT signaling in the pathogenesis of
927 inflammatory bowel disease. *Pharmacol. Res.* 76, 1–8
928 40 Bolton, C. *et al.* (2022) An integrated taxonomy for monogenic inflammatory bowel
929 disease. *Gastroenterology* 162, 859–876
930 41 Hu, M. *et al.* (2022) Fucosyltransferase 2: A genetic risk factor for intestinal diseases.
931 *Front. Microbiol.* 13, 940196

932 42 McGovern, D.P.B. *et al.* (2010) Fucosyltransferase 2 (FUT2) non-secretor status is
933 associated with Crohn's disease. *Hum. Mol. Genet.* 19, 3468–3476
934 43 Nayar, S. *et al.* (2021) A myeloid-stromal niche and gp130 rescue in NOD2-driven
935 Crohn's disease. *Nature* 593, 275–281
936 44 Watanabe, T. *et al.* (2005) NOD2 regulation of Toll-like receptor responses and the
937 pathogenesis of Crohn's disease. *Gut* 54, 1515–1518
938 45 Lesage, S. *et al.* (2002) CARD15/NOD2 mutational analysis and genotype-phenotype
939 correlation in 612 patients with inflammatory bowel disease. *Am. J. Hum. Genet.* 70,
940 845–857
941 46 Hegarty, L.M. *et al.* (2023) Macrophages in intestinal homeostasis and inflammatory
942 bowel disease. *Nat. Rev. Gastroenterol. Hepatol.* 20, 538–553
943 47 West, N.R. *et al.* (2017) Oncostatin M drives intestinal inflammation and predicts
944 response to tumor necrosis factor-neutralizing therapy in patients with inflammatory
945 bowel disease. *Nat. Med.* 23, 579–589
946 48 Timshel, P.N. *et al.* (2020) Genetic mapping of etiologic brain cell types for obesity.
947 *eLife* 9,
948 49 Yengo, L. *et al.* (2018) Meta-analysis of genome-wide association studies for height and
949 body mass index in ~700000 individuals of European ancestry. *Hum. Mol. Genet.* 27,
950 3641–3649
951 50 Lee, J.J. *et al.* (2018) Gene discovery and polygenic prediction from a genome-wide
952 association study of educational attainment in 1.1 million individuals. *Nat. Genet.* 50,
953 1112–1121
954 51 Dennison, T.W. *et al.* (2024) Patient-derived organoid biobank identifies epigenetic
955 dysregulation of intestinal epithelial MHC-I as a novel mechanism in severe Crohn's
956 Disease. *Gut* 73, 1464–1477
957 52 Hundorfean, G. *et al.* (2007) Luminal antigens access late endosomes of intestinal
958 epithelial cells enriched in MHC I and MHC II molecules: in vivo study in Crohn's ileitis.
959 *Am. J. Physiol. Gastrointest. Liver Physiol.* 293, G798-808
960 53 Linke, A. *et al.* (2022) Antigen Cross-Presentation by Murine Proximal Tubular Epithelial
961 Cells Induces Cytotoxic and Inflammatory CD8+ T Cells. *Cells* 11,
962 54 Limmer, A. *et al.* (2000) Efficient presentation of exogenous antigen by liver endothelial
963 cells to CD8+ T cells results in antigen-specific T-cell tolerance. *Nat. Med.* 6, 1348–1354
964 55 Malik, A. *et al.* (2023) Epithelial IFNγ signalling and compartmentalized antigen
965 presentation orchestrate gut immunity. *Nature* 623, 1044–1052
966 56 Na, Y.R. *et al.* (2019) Macrophages in intestinal inflammation and resolution: a potential
967 therapeutic target in IBD. *Nat. Rev. Gastroenterol. Hepatol.* 16, 531–543
968 57 Bain, C.C. *et al.* (2013) Resident and pro-inflammatory macrophages in the colon
969 represent alternative context-dependent fates of the same Ly6Chi monocyte precursors.
970 *Mucosal Immunol.* 6, 498–510
971 58 Kamada, N. *et al.* (2008) Unique CD14 intestinal macrophages contribute to the
972 pathogenesis of Crohn disease via IL-23/IFN-gamma axis. *J. Clin. Invest.* 118, 2269–
973 2280
974 59 Thiesen, S. *et al.* (2014) CD14(hi)HLA-DR(dim) macrophages, with a resemblance to
975 classical blood monocytes, dominate inflamed mucosa in Crohn's disease. *J. Leukoc.*
976 *Biol.* 95, 531–541
977 60 Hedl, M. *et al.* (2016) JAK2 Disease-Risk Variants Are Gain of Function and JAK
978 Signaling Threshold Determines Innate Receptor-Induced Proinflammatory Cytokine
979 Secretion in Macrophages. *J. Immunol.* 197, 3695–3704

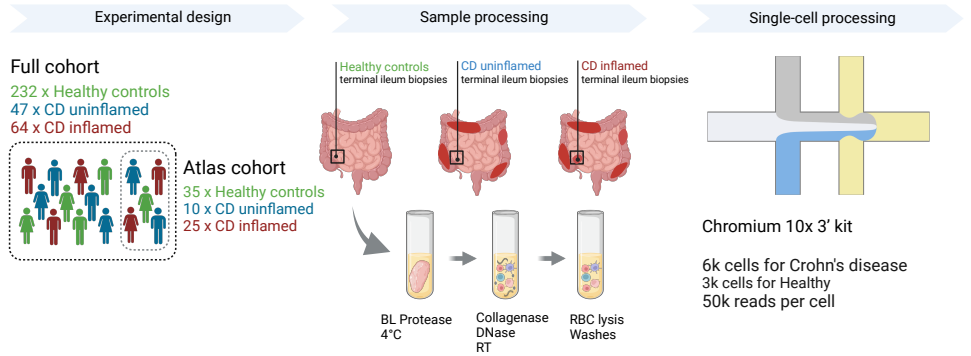
61 Park, C. *et al.* (2000) Immune response in Stat2 knockout mice. *Immunity* 13, 795–804
62 Desai, H.R. *et al.* (2017) Macrophage JAK2 deficiency protects against high-fat diet-induced inflammation. *Sci. Rep.* 7, 7653
63 Jagadeesh, K.A. *et al.* (2022) Identifying disease-critical cell types and cellular processes by integrating single-cell RNA-sequencing and human genetics. *Nat. Genet.* 54, 1479–1492
64 Beaudoin, M. *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
65 Franke, A. *et al.* (2010) Genome-wide meta-analysis increases to 71 the number of confirmed Crohn's disease susceptibility loci. *Nat. Genet.* 42, 1118–1125
66 UK IBD Genetics Consortium *et al.* (2009) Genome-wide association study of ulcerative colitis identifies three new susceptibility loci, including the HNF4A region. *Nat. Genet.* 41, 1330–1334
67 Fleming, S.J. *et al.* (2019) Unsupervised removal of systematic background noise from droplet-based single-cell experiments using CellBender. *BioRxiv* DOI: 10.1101/791699
68 Lun, A.T.L. *et al.* (2019) EmptyDrops: distinguishing cells from empty droplets in droplet-based single-cell RNA sequencing data. *Genome Biol.* 20, 63
69 Wolock, S.L. *et al.* (2019) Scrublet: Computational Identification of Cell Doublets in Single-Cell Transcriptomic Data. *Cell Syst.* 8, 281-291.e9
70 van der Walt, S. *et al.* (2014) scikit-image: image processing in Python. *PeerJ* 2, e453
71 Luecken, M.D. and Theis, F.J. (2019) Current best practices in single-cell RNA-seq analysis: a tutorial. *Mol. Syst. Biol.* 15, e8746
72 Wolf, F.A. *et al.* (2018) SCANPY: large-scale single-cell gene expression data analysis. *Genome Biol.* 19, 15
73 Cattell, R.B. (1966) The scree test for the number of factors. *Multivariate Behav. Res.* 1, 245–276
74 Satopaa, V. *et al.* (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
75 Polański, K. *et al.* (2020) BBKNN: fast batch alignment of single cell transcriptomes. *Bioinformatics* 36, 964–965
76 Traag, V.A. *et al.* (2019) From Louvain to Leiden: guaranteeing well-connected communities. *Sci. Rep.* 9, 5233
77 Chicco, D. and Jurman, G. (2020) The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genomics* 21, 6
78 Schelker, M. *et al.* (2017) Estimation of immune cell content in tumour tissue using single-cell RNA-seq data. *Nat. Commun.* 8, 2032
79 Domínguez Conde, C. *et al.* (2022) Cross-tissue immune cell analysis reveals tissue-specific features in humans. *Science* 376, eabl5197
80 Finak, G. *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
81 Zimmerman, K.D. *et al.* (2021) A practical solution to pseudoreplication bias in single-cell studies. *Nat. Commun.* 12, 738
82 Benjamini, Y. and Hochberg, Y. (1995) Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society.*

1028 *Series B (Methodological)* 57, 289–300
1029 83 Korotkevich, G. *et al.* (2016) Fast gene set enrichment analysis. *BioRxiv* DOI:
1030 10.1101/060012
1031 84 Fabregat, A. *et al.* (2017) Reactome pathway analysis: a high-performance in-memory
1032 approach. *BMC Bioinformatics* 18, 142
1033 85 Subramanian, A. *et al.* (2005) Gene set enrichment analysis: a knowledge-based
1034 approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci USA*
1035 102, 15545–15550
1036 86 Zeisel, A. *et al.* (2018) Molecular architecture of the mouse nervous system. *Cell* 174,
1037 999-1014.e22
1038 87 Skene, N.G. *et al.* (2018) Genetic identification of brain cell types underlying
1039 schizophrenia. *Nat. Genet.* 50, 825–833
1040 88 Dougherty, J.D. *et al.* (2010) Analytical approaches to RNA profiling data for the
1041 identification of genes enriched in specific cells. *Nucleic Acids Res.* 38, 4218–4230

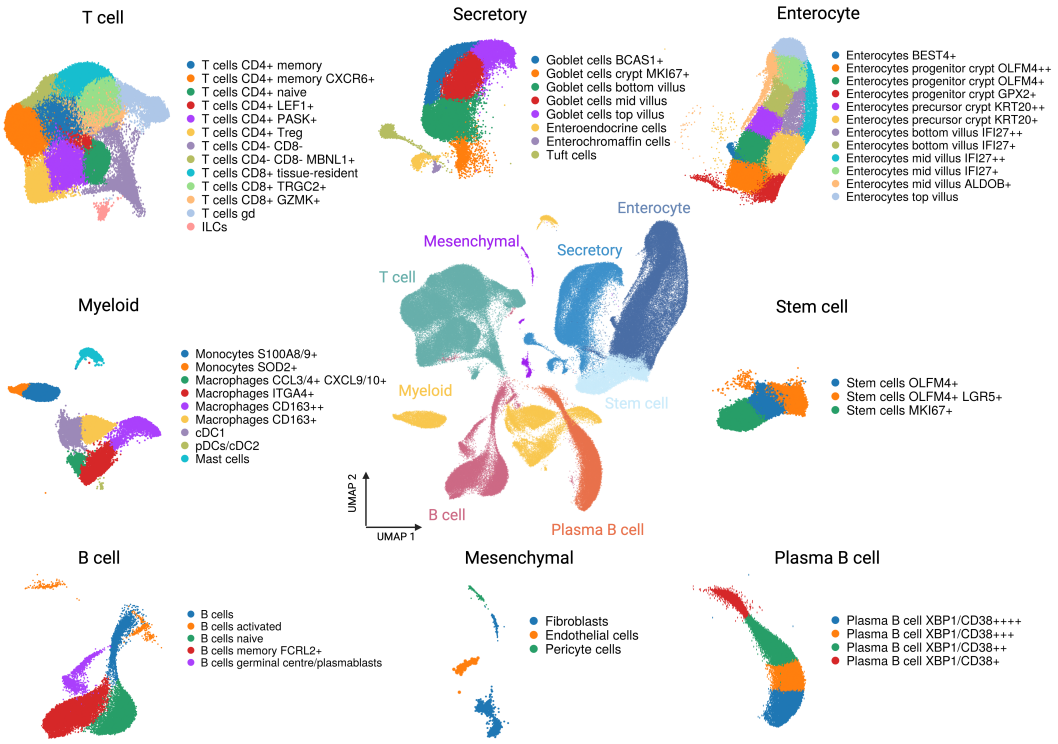
Figures

Fig. 1. Single-cell expression atlas of the terminal ileum in healthy controls and patients with Crohn's Disease.

(a) Experimental design and sample processing

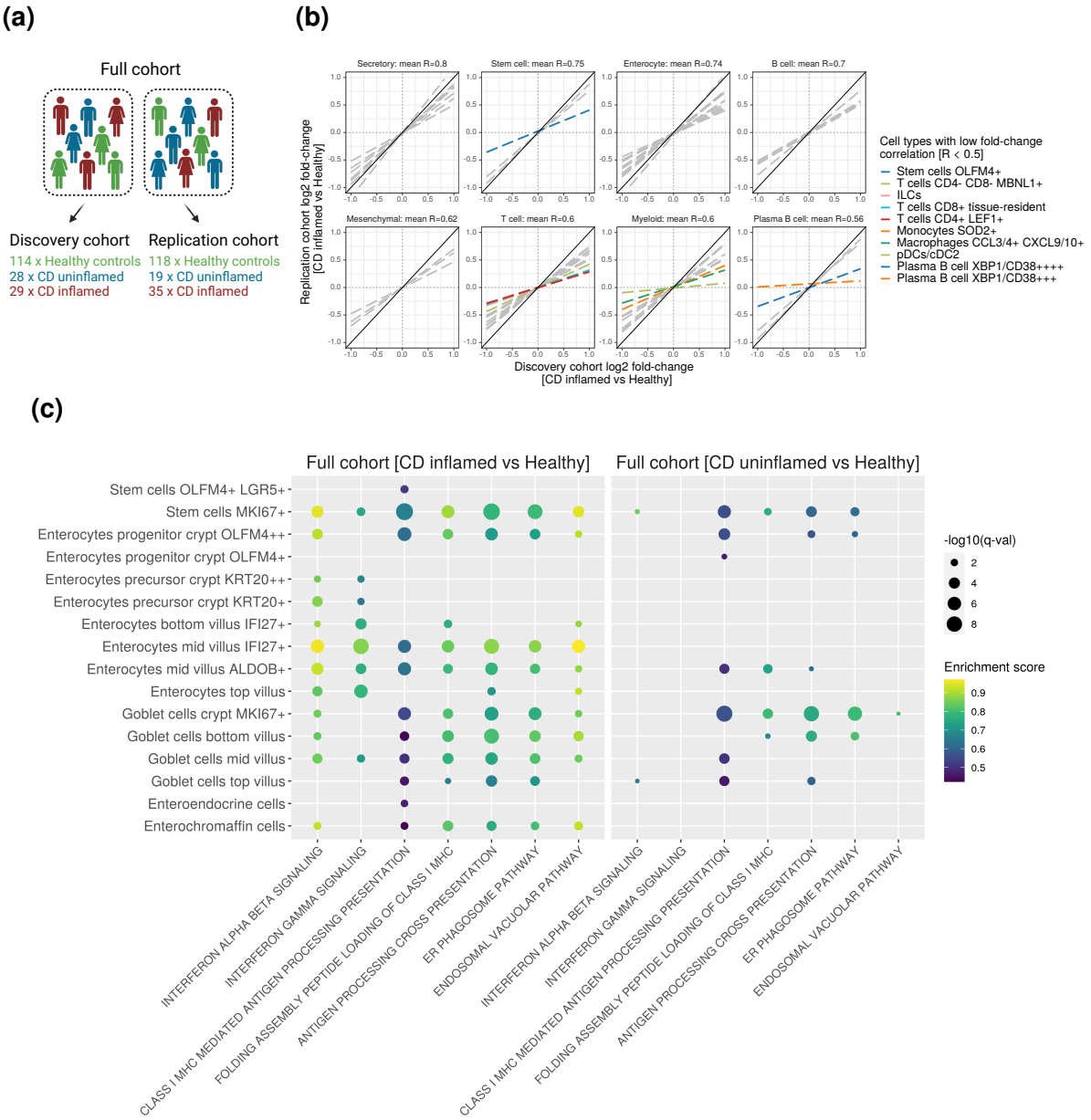


(b) UMAP projection of cells from the Atlas cohort



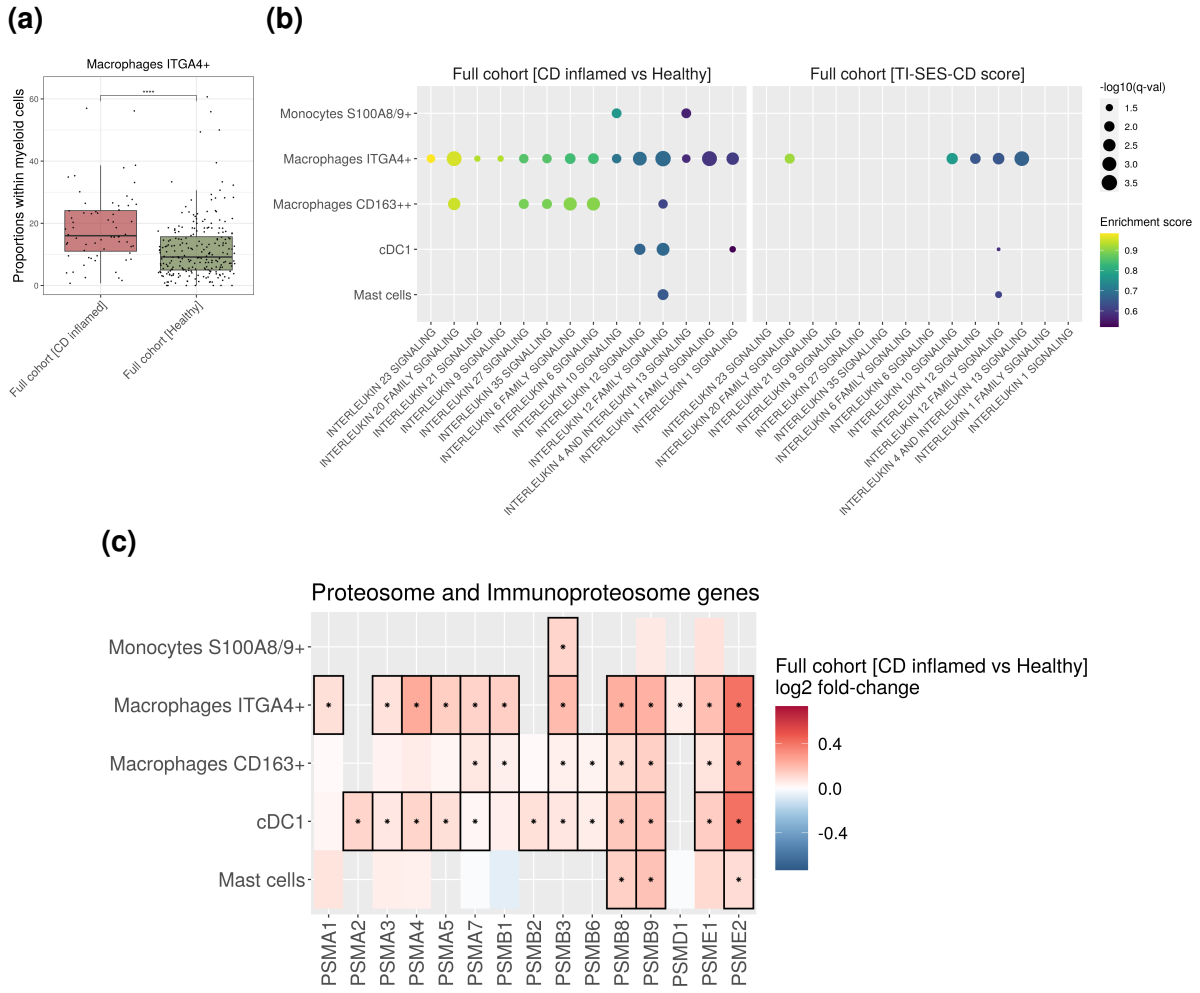
(a) Terminal ileum biopsies from 343 individuals were dissociated to single cells on ice in Hank's balanced salt solution (HBSS) containing Bacillus Licheniformis (BL) protease. This was followed by a brief incubation in collagenase and then red blood cell (RBC) lysis buffer. Single cell suspensions were then profiled with the Chromium 10X 3' kit (Methods). **(b)** Centre: Uniform Manifold Approximation and Projection (UMAP) of ∼216k cells from the atlas cohort that meet quality control criteria (Methods), with eight colors representing the primary cell populations. Surrounding the central UMAP, these major populations are further subdivided into 57 distinct cell subtypes.

Fig. 2. Replicable differential gene expression signatures in CD epithelial cells enriched for interferon signaling and MHC I antigen presentation.



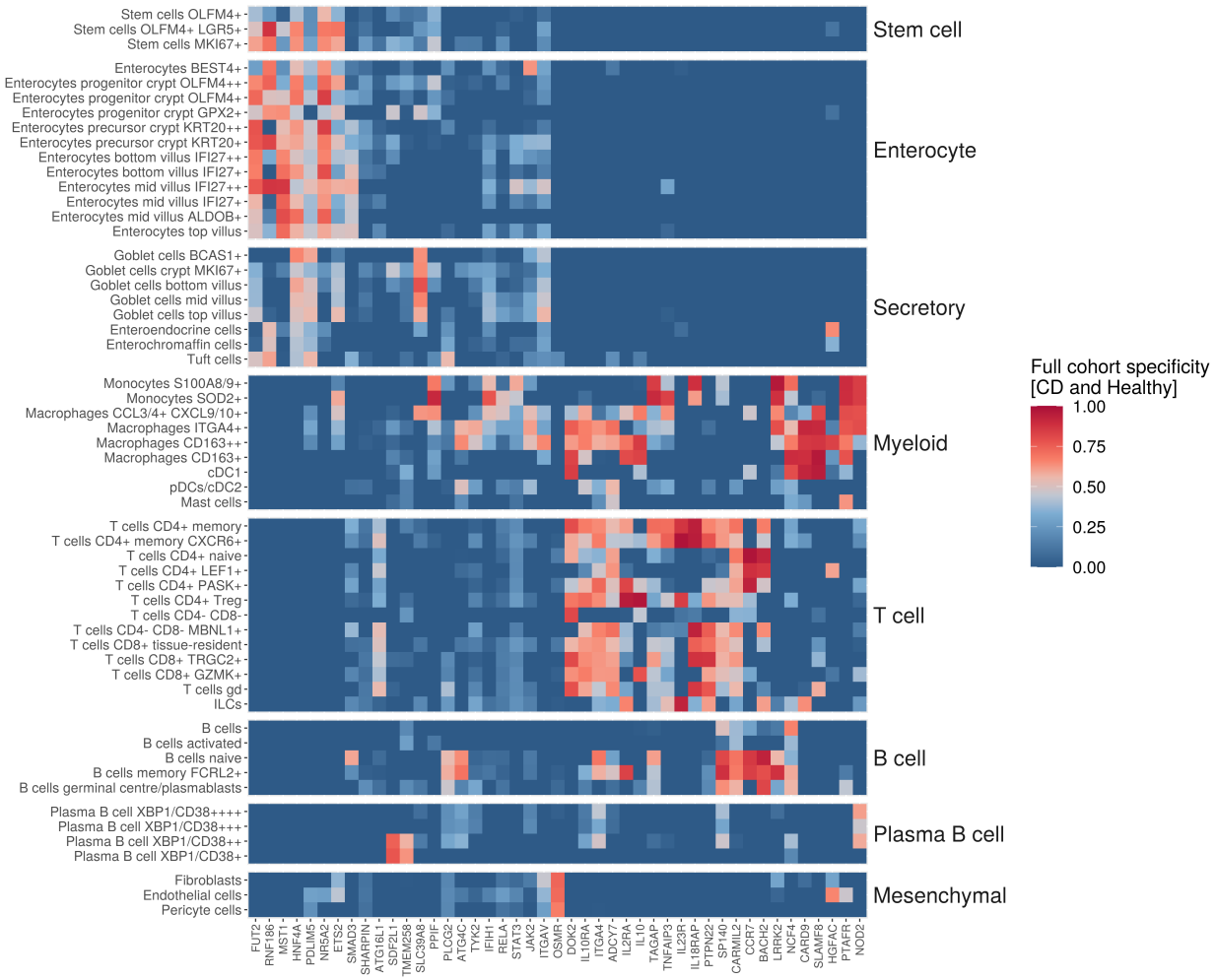
(a) Design of discovery and replication cohorts; i) first, 70 samples in the atlasing cohort were equally distributed at random between discovery and replication cohorts ii) the remaining 273 samples not included in the atlasing cohort were then randomly assigned to either cohort **(b)** Linear regression between log2 fold changes of differentilly expressed genes (without thresholding, as outlined in Methods) in the discovery (x-axis) and replication (y-axis) datasets. The reported mean correlation represents the average coefficient across the major cell populations. Highlighted cell types with low fold-change correlation ($R < 0.5$, Table S5) **(c)** Gene set enrichment analysis of differential gene expression z-scores from the full cohort. Results for epithelial cell types with high fold-change correlation ($R \geq 0.5$).

Fig. 3. *ITGA4*-positive macrophages upregulate cytokine and proteasome genes during inflammation.



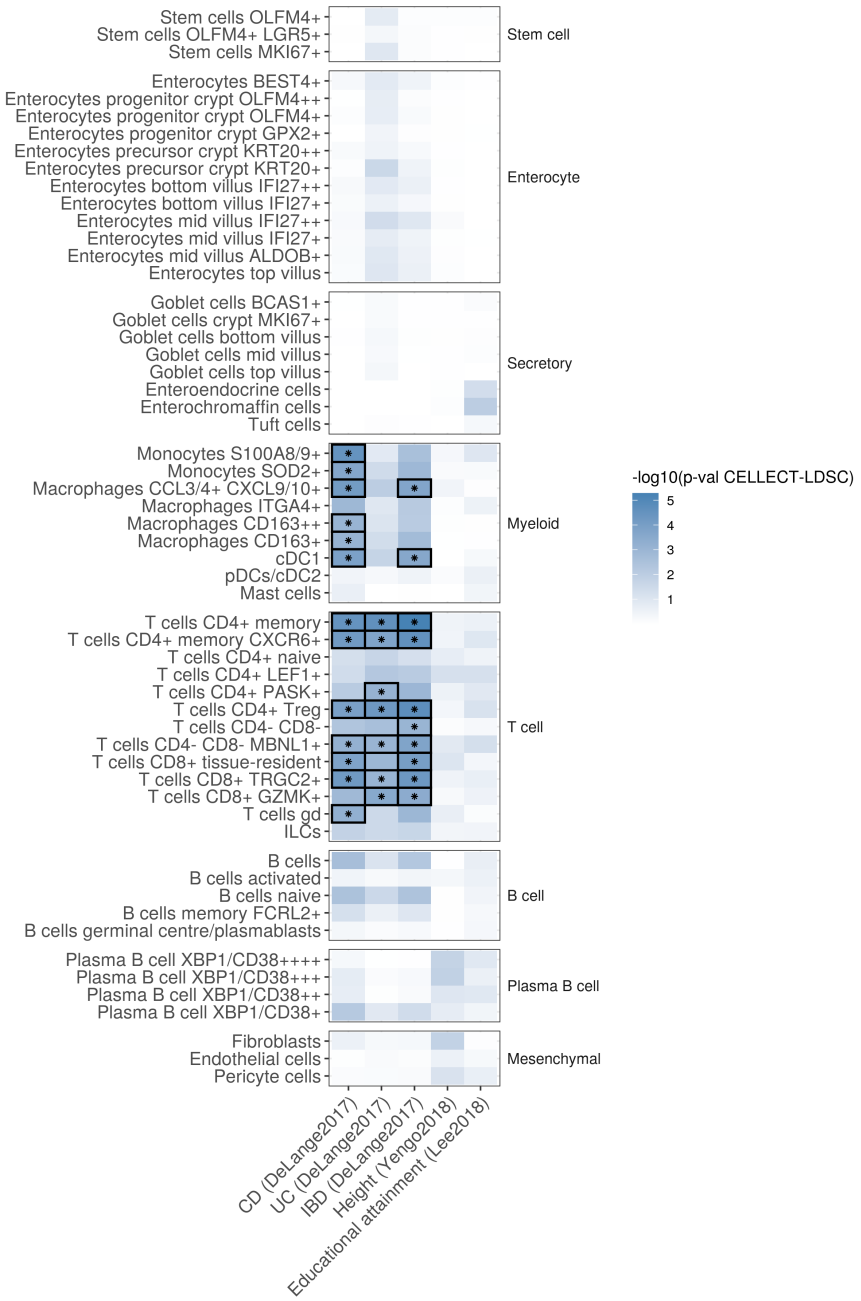
(a) Proportions of *ITGA4*-positive macrophages within myeloid cells across inflamed CD and healthy samples (adjusted p values **** < 0.0001, t-test). **(b)** Gene set enrichment analysis of differential gene expression between inflamed CD and healthy samples and across CD samples with varying severity of inflammation- CD samples were stratified based on the simple endoscopic score applied to terminal ileal segment (TI-SES-CD). Enrichment shown for replicable myeloid cell types ($R \geq 0.5$, Table S5) **(c)** Fold change and significance of proteasome and immunoproteasome genes upregulated in the replicable myeloid cell populations.

Fig. 4. IBD risk genes show lineage-specific expression across terminal ileum cell types.



Specificity of IBD risk gene expression across 57 terminal ileum cell-types in the full cohort. For prioritisation criteria, see Methods.

Fig. 5. Myeloid and T cells are enriched for CD and UC heritability.



Significance of CD, UC and IBD heritability in specifically expressed genes of each cell-type (family-wise error rate (FWER) < 0.05). Height and Educational Attainment were used as negative controls.

Tables

Table 1. Demographics of healthy and CD patients in the IBDverse.

	Healthy (%)	Crohn's Disease (%)	P-value
N	232	111	
Inflamed	0 (0.0)	64 (57.7)	$<$0.001
TI-SES-CD			
$[0-3)$	232 (100.0)	47 (42.3)	
$[3-6)$	0 (0.0)	50 (45.0)	
$[6-9)$	0 (0.0)	14 (12.6)	
Sex = M	118 (50.9)	48 (43.2)	0.205
Smoking Status			0.208
No	178 (76.7)	77 (69.4)	
Yes	25 (10.8)	21 (18.9)	
Ex-Smoker	25 (10.8)	12 (10.8)	
Vape Smoker	4 (1.7)	1 (0.9)	
Mean Age (SD)	49.27 (13.31)	40.87 (12.12)	$<$0.001

Absolute number and proportions of demographics across IBDverse. Inflamation was determined based on TI-SES-CD = TI endoscopic score for CD, an aggregated score quantifying the degree of inflammation at several points in the TI. The values of the TI-SES-CD range from 0 to 9 (9 indicating the most severe inflammation). Patients were stratified into two groups: inflamed TI-SES-CD $\geq$ 3 and uninflamed TI-SES-CD $<$ 3. A significant difference in mean age was observed between healthy controls and CD patients (p $<$ 0.001, t-test).