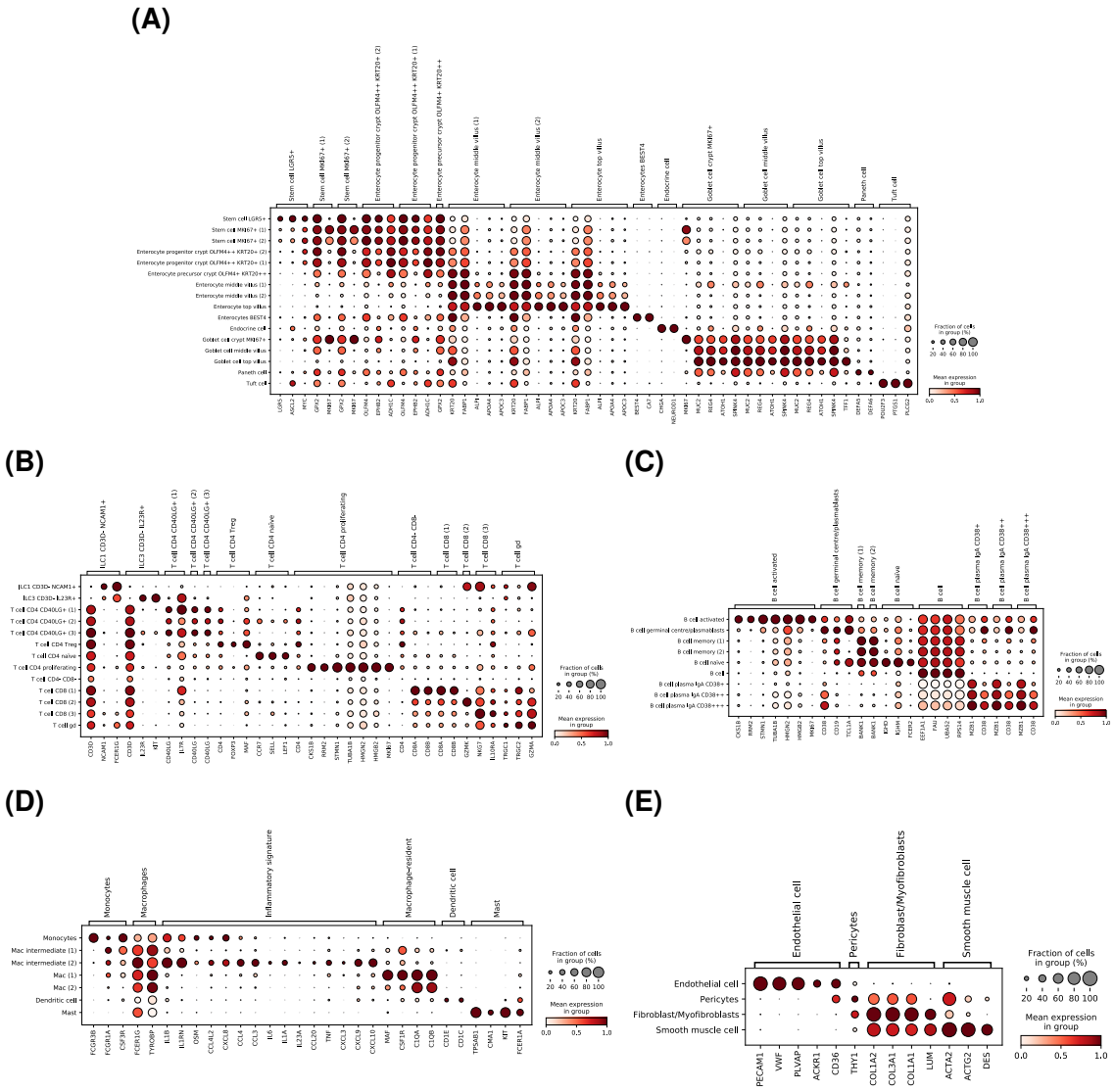
Supplementary Figures



(F) Specifically expressed genes are robust across discovery and replication cohorts.

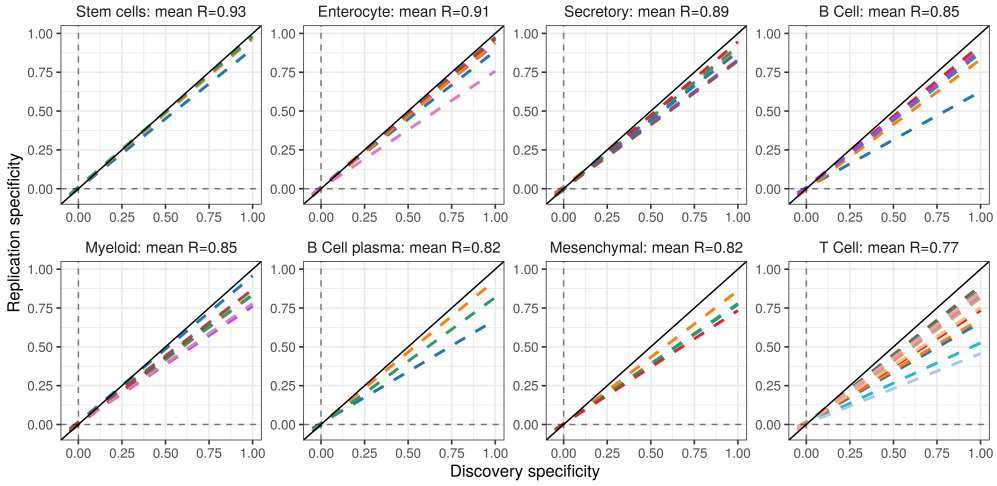
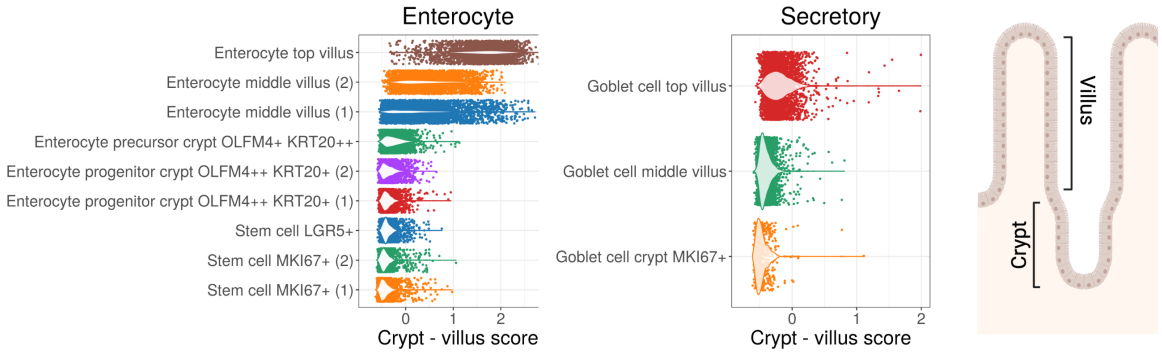


Figure S1

Figure S1. Marker gene expression used to curate terminal ileum atlas cluster annotations.
(A-E) Transcriptional signatures of literature or expert-curated markers that distinguish epithelial, T cell, B cell, myeloid and mesenchymal cell types. Dot size is proportional to the percentage of cells expressing a given marker within a cluster and the colour intensity denotes average gene expression within the cluster. (F) For each of the 49 cell types we fitted linear regression model between computationally-determined specifically expressed genes (no thresholds applied, Methods) in the discovery (x axis) and replication (y axis) datasets. For colour legend see Figure 1B.

(A) Enterocyte and secretory cells scored for crypt-villus signature genes.



(B) Developmental trajectories of the epithelial cell types.

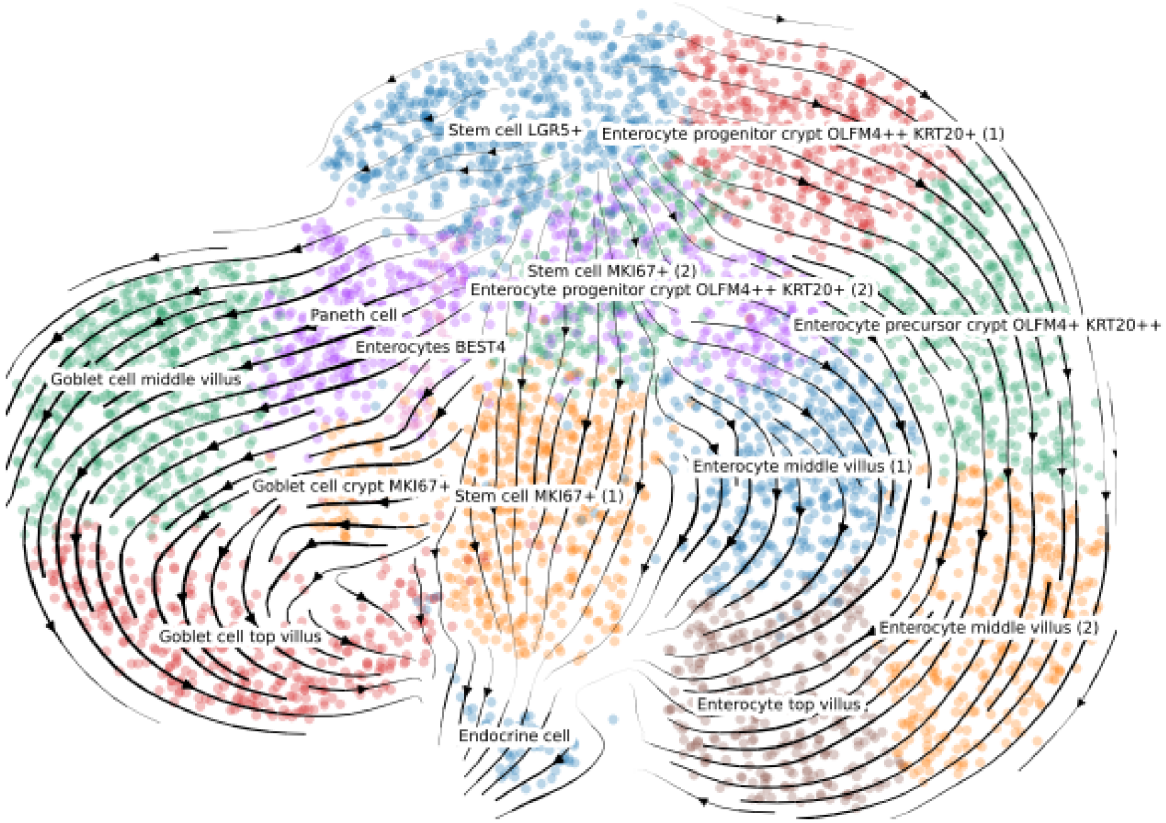


Figure S2

Figure S2. Epithelial cell types represent the crypt-villus axis differentiation. (A) We identified multiple subtypes of enterocytes and goblet cells. To distinguish between them, we used two gene signatures from (Moor et al. 2018) to position epithelial and goblet cells along the crypt-villus axis. A subpopulation of enterocytes, that we defined as "enterocyte top villus" showed higher score (higher gene expression) for the gene signature (including *APOA4, APOC3, ALPI* genes), which defines cells at the tip of the intestinal villi. Enterocyte middle villus cells, progenitors or stem cells presented a lower expression of those genes. Similarly, we scored goblet cells for top-villus gene signature including *EGFR, KLF4 and NT5E* genes. Goblet top villus cells presented slightly higher score than goblet middle villus cells and goblet cells at the crypt base. Non crypt-villus migrating cells such as tuft, paneth and endocrine cells, were excluded from the plot. (B) To infer the transition of epithelial cells from the bottom crypt to the top villus we used RNA velocity analysis (Methods) on all epithelial cell types except tuft cells. The figure shows UMAP of epithelial cells from discovery terminal ileum atlas (both Crohn's disease and healthy individuals) with arrows representing estimated velocity vectors i.e. transitions between cell states. We observed two main lineages of cells that arised from stem cells and transitioned into absorptive enterocytes (enterocyte top villus) and secretory cells (goblet cell top villus)

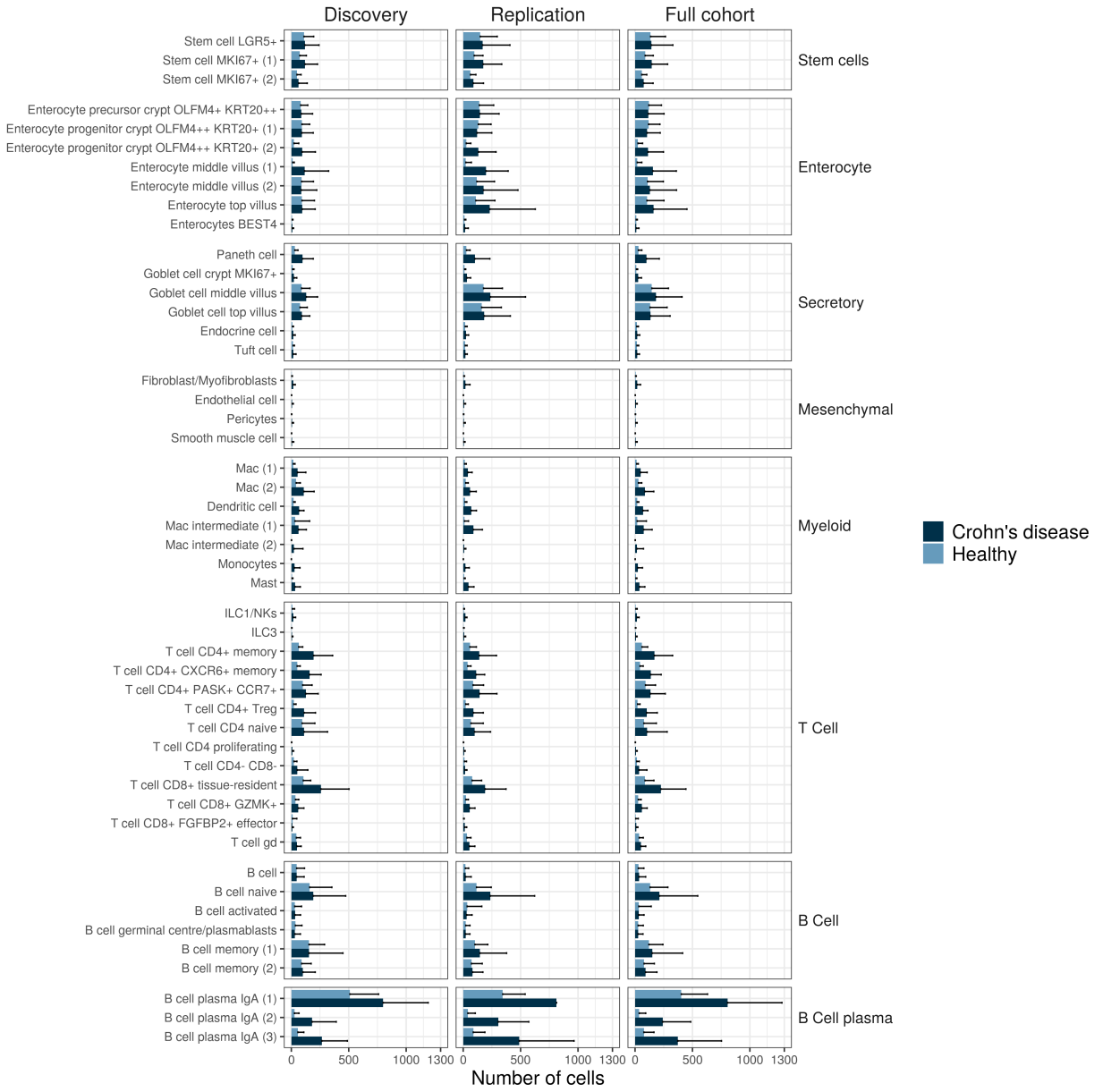
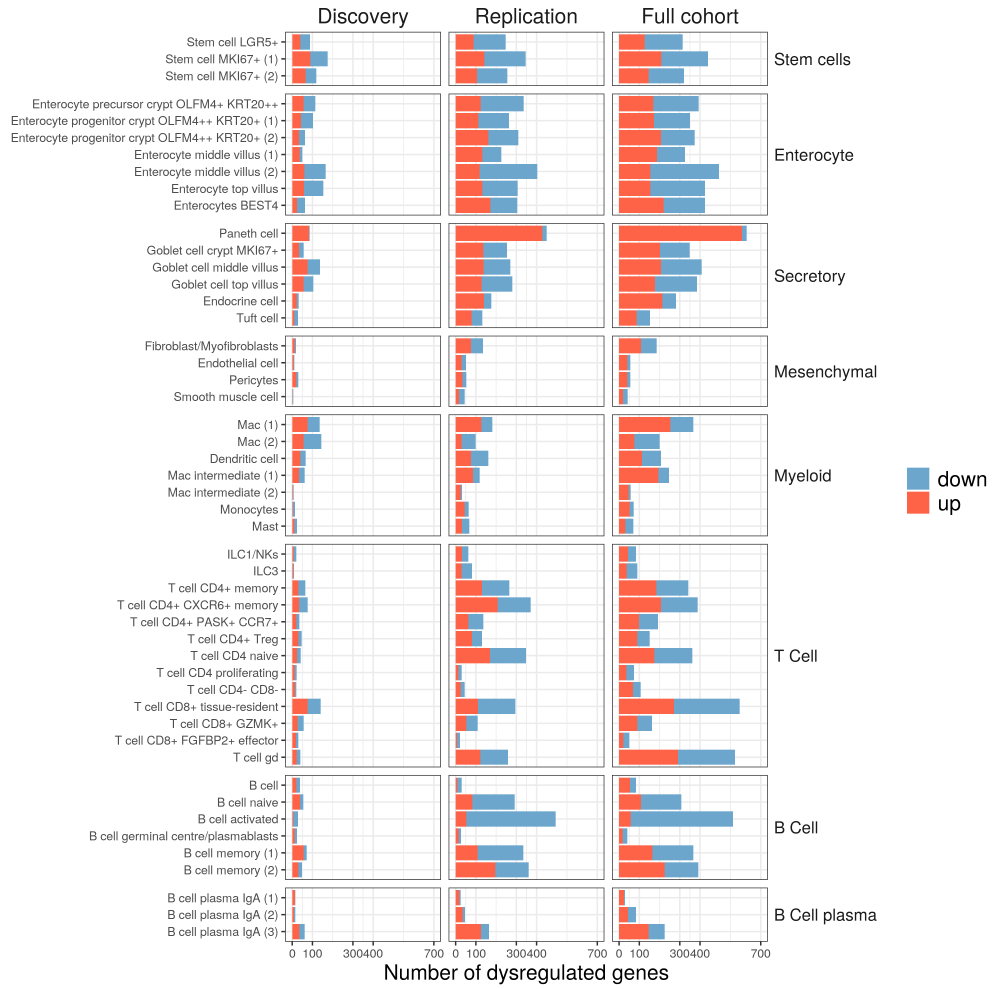

Discovery
Replication
Full cohort
Stem cell LGR5+
Stem cell MKI67+ (1)
Stem cell MKI67+ (2)
Stem cells
Enterocyte precursor crypt OLFM4+ KRT20++
Enterocyte progenitor crypt OLFM4++ KRT20+ (1)
Enterocyte progenitor crypt OLFM4++ KRT20+ (2)
Enterocyte middle villus (1)
Enterocyte middle villus (2)
Enterocyte top villus
Enterocytes BEST4
Enterocyte
Paneth cell
Goblet cell crypt MKI67+
Goblet cell middle villus
Goblet cell top villus
Endocrine cell
Tuft cell
Secretory
Fibroblast/Myofibroblasts
Endothelial cell
Pericytes
Smooth muscle cell
Mesenchymal
Mac (1)
Mac (2)
Dendritic cell
Mac intermediate (1)
Mac intermediate (2)
Monocytes
Mast
Myeloid
ILC1/NKs
ILC3
T cell CD4+ memory
T cell CD4+ CXCR6+ memory
T cell CD4+ PASK+ CCR7+
T cell CD4+ Treg
T cell CD4 naive
T cell CD4 proliferating
T cell CD4- CD8-
T cell CD8+ tissue-resident
T cell CD8+ GZMK+
T cell CD8+ FGFBP2+ effector
T cell gd
T Cell
B cell
B cell naive
B cell activated
B cell germinal centre/plasmablasts
B cell memory (1)
B cell memory (2)
B Cell
B cell plasma IgA (1)
B cell plasma IgA (2)
B cell plasma IgA (3)
B Cell plasma
0
500
1000
1300
Number of cells
Crohn's disease
Healthy

Figure S3

Figure S3. Cellular composition of terminal ileum atlas.
Mean number of cells in each of 49 cell types across healthy (light blue) and Crohn's disease (dark blue) samples from discovery, replication and full cohorts. Error bars represent standard deviation from the mean number of cells.

(A) Frequency of differentially expressed genes.



(B) Number of dysregulated genes is highly correlated with cell abundance.

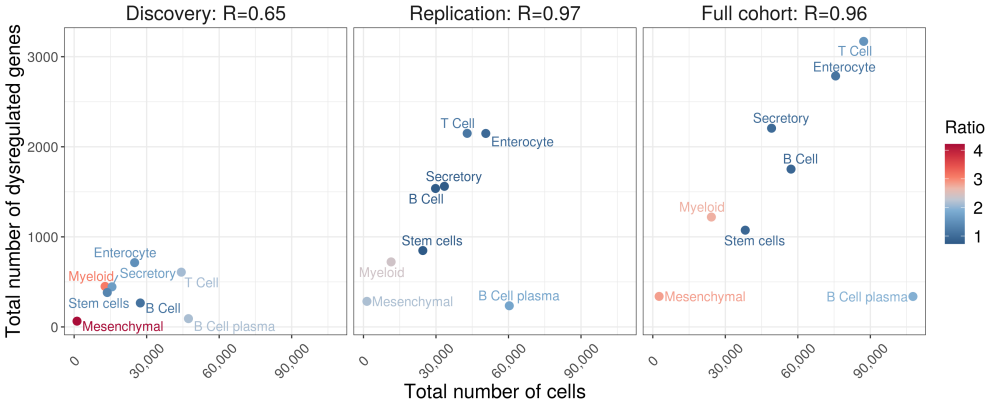
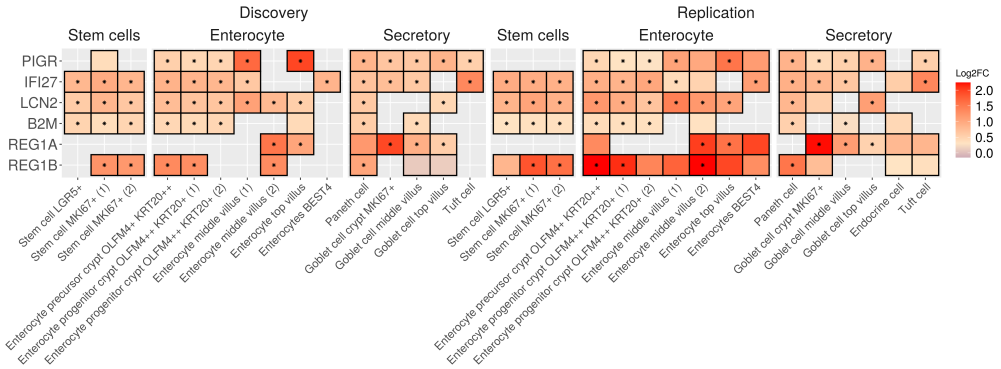


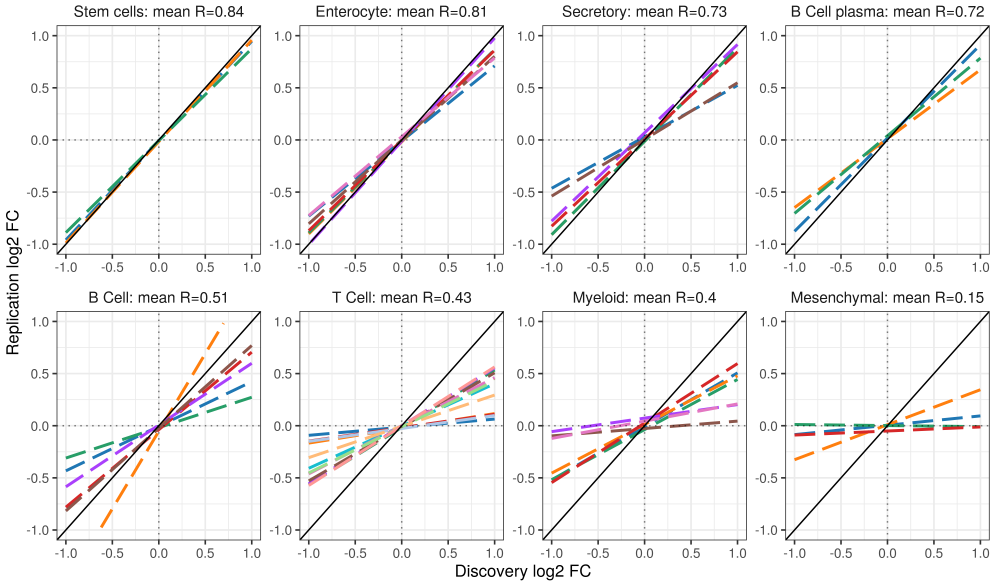
Figure S4

Figure S4. Differentially expressed genes in Crohn's disease (CD) across all 49 cell types.
(A) Number of significantly (FDR $<$ 5%) up- and down-regulated genes (x axis) found in each of 49 cell types (y axis) across discovery, replication and full cohorts. (B) Shown are the total number of cells in each major cell population (x axis), the total number of significantly dysregulated genes (FDR $<$ 5%) and ratio of cells in CD vs healthy (colour) across discovery, replication and full cohorts. Pearson correlation coefficient (R) was calculated for all major cell populations except outlying B plasma cells.

(A) Up-regulated genes in CD that are highly concordant across discovery and replication cohorts.



(B) Concordance of differential gene expression across discovery and replication cohorts.



(C) Correlations between DEG and SEG reinforces the modular structure of terminal ileum cell type transcriptomes.

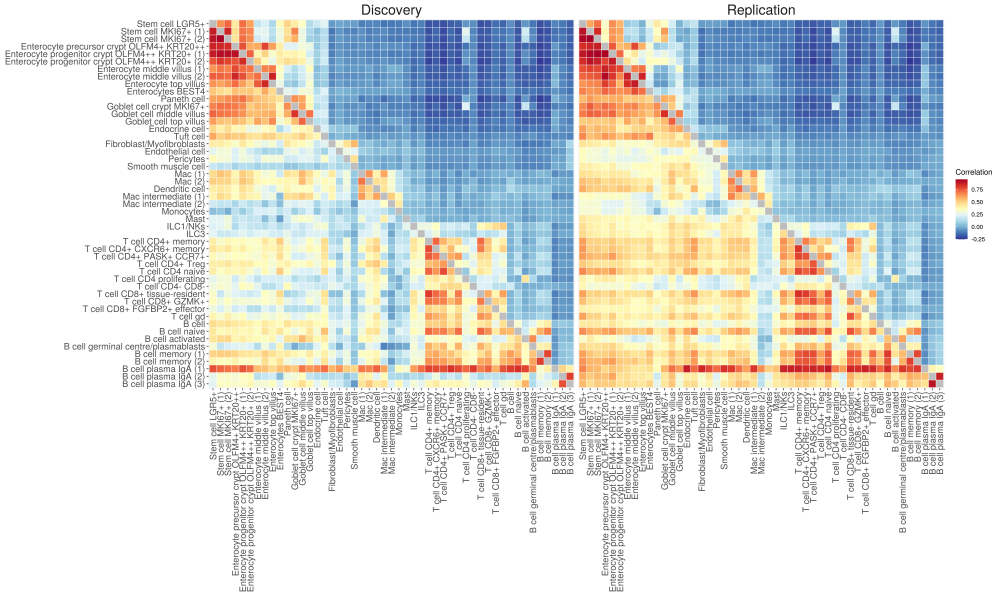
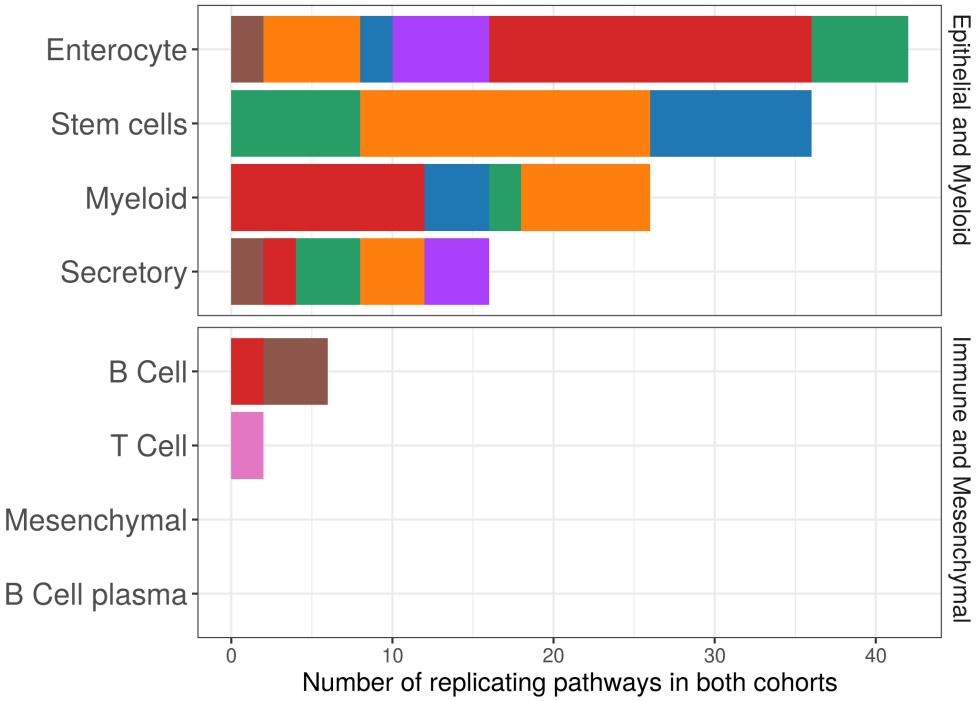


Figure S5

Figure S5. Reproducibility of differentially expressed genes across discovery and replication cohorts.
(A) DGE log2 fold changes of set of genes significantly up-reguated (FDR $<$ 5%) across epithelial cells in the discovery and replication cohorts. Genes dysregulated in both cohorts in the same cell types (so-called replicable) were denoted by an asterisk. (B) For each of the 49 cell types we fitted linear regression model between DEGs log2 fold changes in the discovery (x axis) and replication (y axis) datasets. For colour legend see Figure 1B. (C) Pearson correlation (across both cohorts) between all 49 cell types in terms of i) upper triangle - specificity of genes and ii) bottom triangle - DGE log2 fold changes. Prior to calculating correlation on DGE log2 fold changes we removed a set of genes with high expression ie. immunoglobulin, mitochondrial and ribosomal.

(A) Number of dysregulated pathways that are replicated across both cohorts.



(B) Immune-related pathways up-regulated in Crohn's disease.

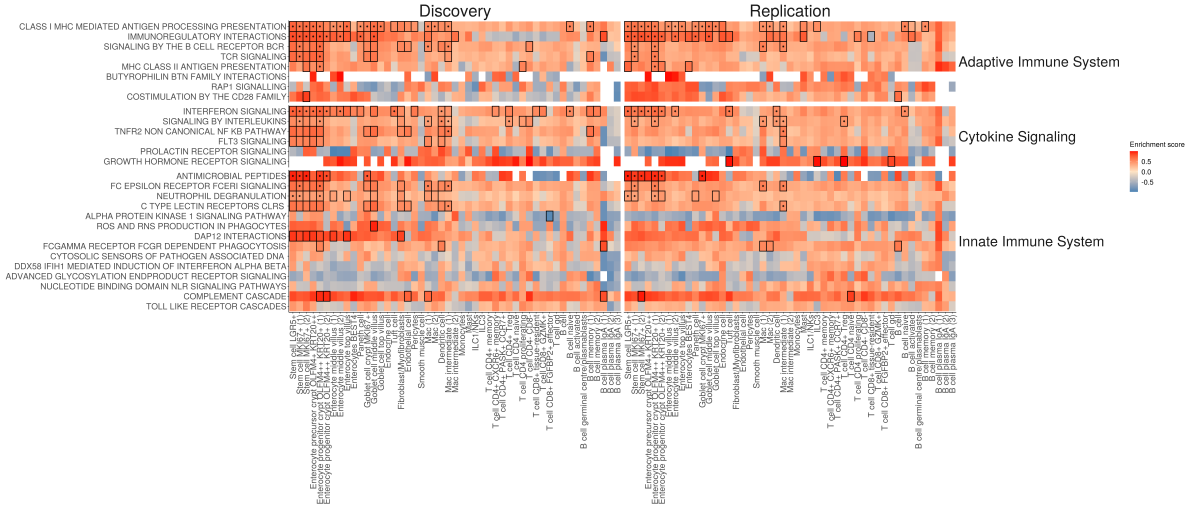
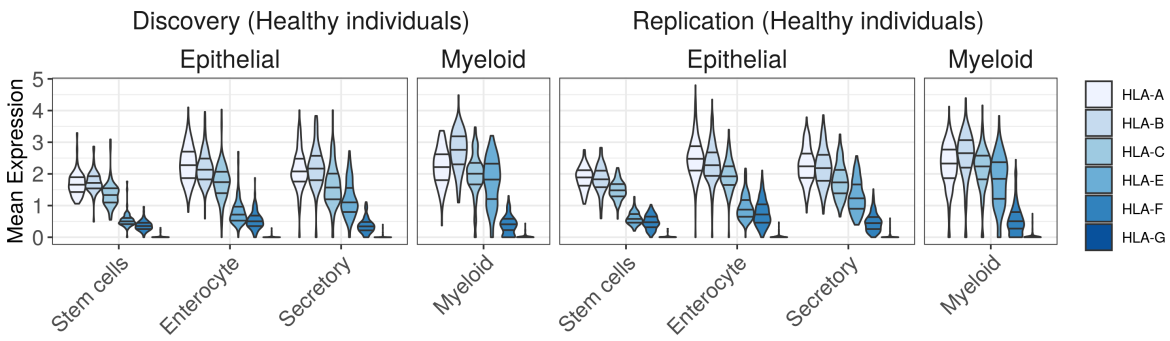


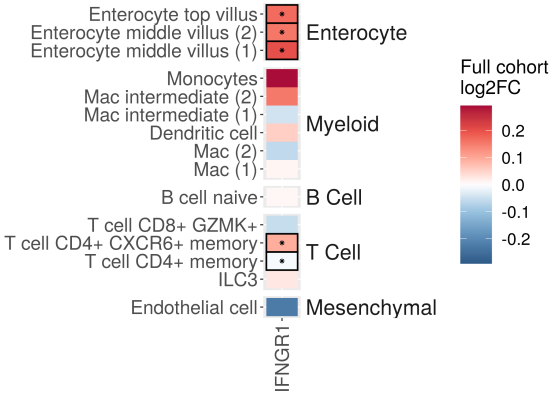
Figure S6

Figure S6. Pathways dysregulated between health and disease.
(A) Number of replicating GSEA pathways that were significantly dysregulated in both cohort cell types. For colour legend see Figure 1B. (B) Enrichment of all cell types across three pathway categories: ”Adaptive immune system”, ”Cytokine signaling”, ”Innate immune system” and their respective sub-pathways. Tiles represent significantly dysregulted pathways (FDR $<$ 5%) in CD versus healthy samples with their enrichment score (colour). Dysregulated pathways replicable in both cohorts were denoted with an asterisk. Pathways are ordered from those most abundantly enriched to those least abundantly enriched.

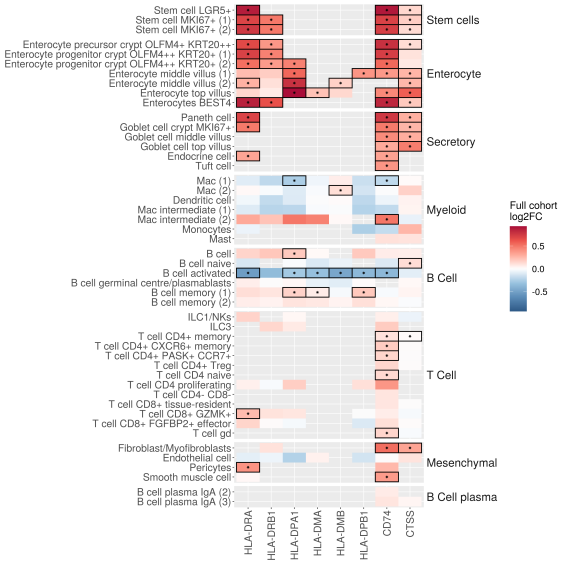
(A) Expression of MHC class I genes



(B) Dysregulation of *IFNGR1* receptor



(C) MHC-II components upregulated in CD epithelial cells



(D) Expression of CD58/CD2 ligand-receptor pair

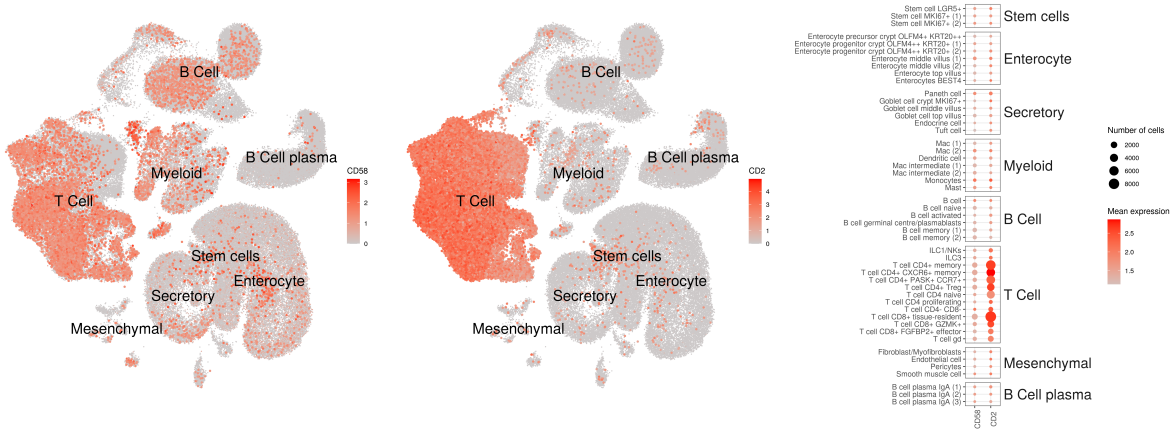
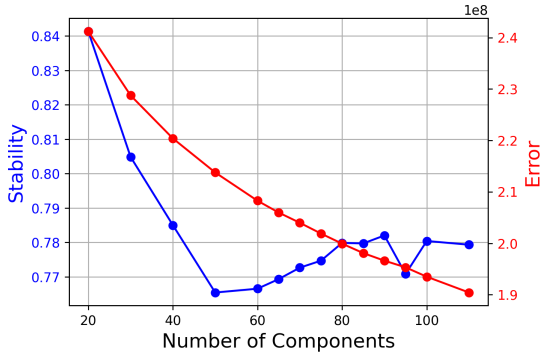


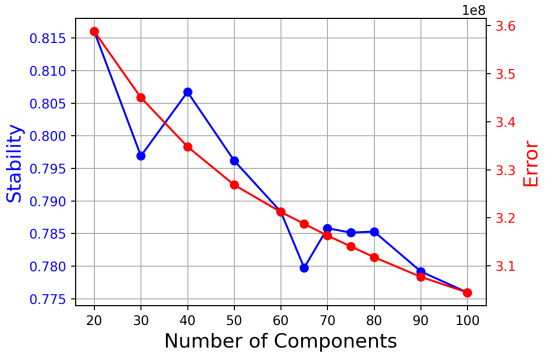
Figure S7

Figure S7. Upregulation of Major Histocompatibility genes and related receptors. (A) We observed consistent expression levels of MHC-I genes across epithelial and myeloid cells in healthy individuals of the discovery and replication cohorts. (B) We showed that *IFNGR1* receptor is upregulated in the full cohort across enterocyte middle and top villus cells during CD. Significant DE genes (FDR $<$ 5%) were denoted by an asterisk. (C) Log2 fold change of MHC-II genes across all cell types. Significant DE genes (FDR $<$ 5%) were denoted by an asterisk. (D) Terminal ileum atlas UMAP projections (from discovery dataset only) coloured by the log1p CP10k expression of *CD58* and *CD2* genes. Alongside dotplot representing the number of cells expressing *CD58* and *CD2* with log1p CP10k $>$ 1.

(A) Optimisation over number of factors in the discovery cohort



(B) Optimisation over number of factors in the replication cohort



(C) Correlation between identified non-negative matrix factorisation (NMF) factors in two cohorts.

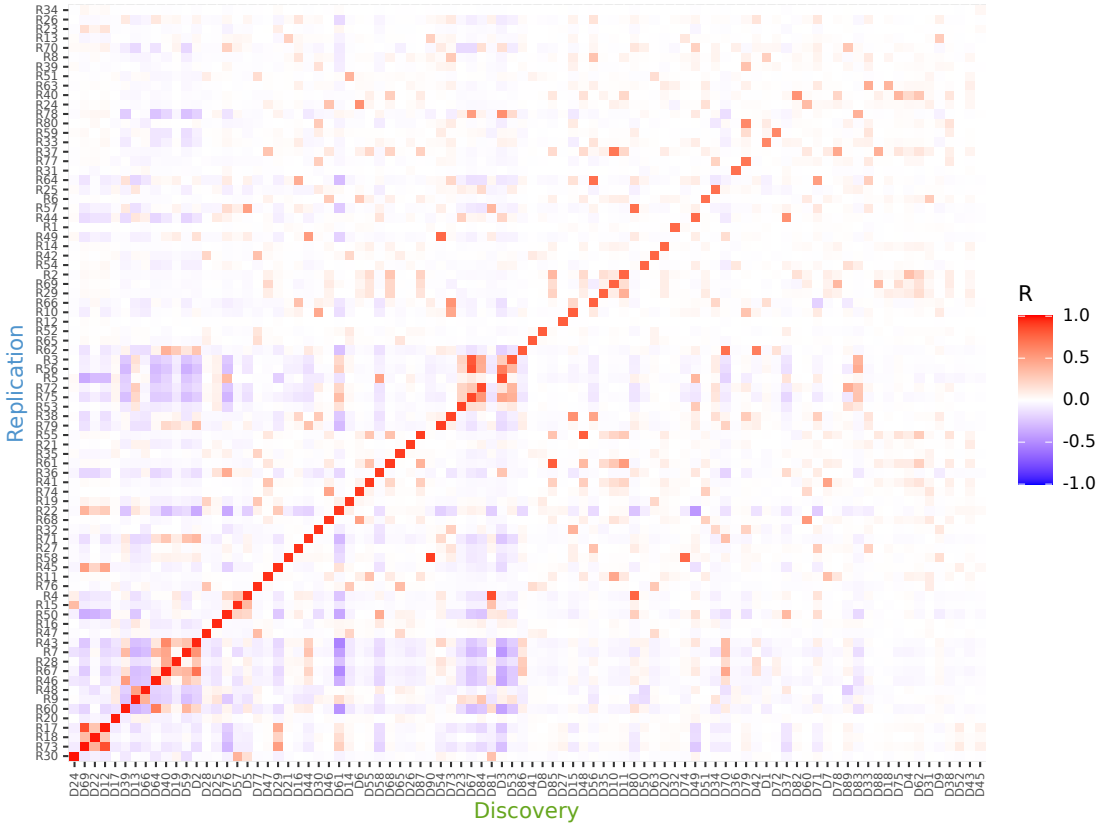
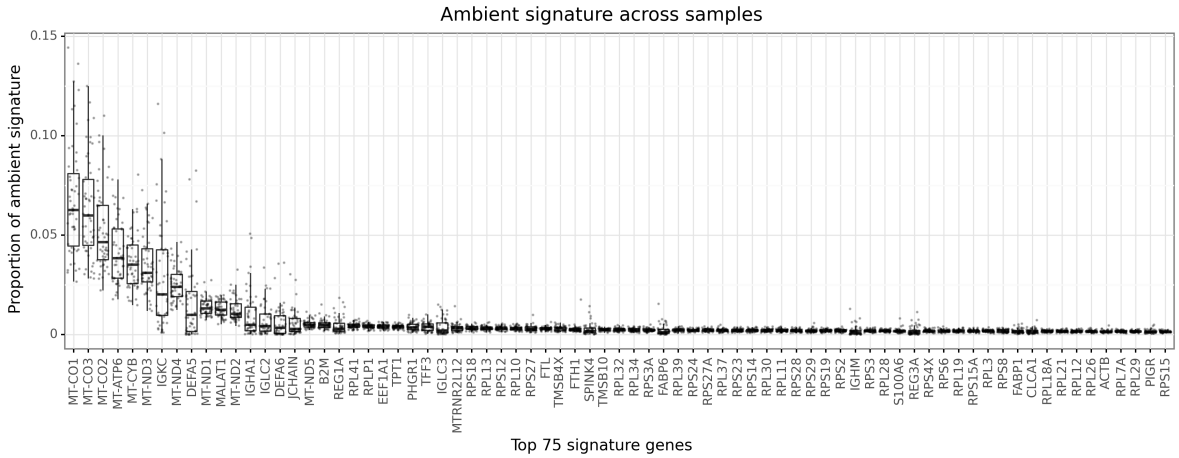


Figure S8

Figure S8. Optimisation of non-negative matrix factorisation (NMF) parameter selection.
Stability (in blue) and error (in red) across iteractions when detecting NMF factors for (A) discovery and (B) replication cohorts showing optima at 90 and 80 factors, respectively. (C) The majority of identified NMF factors were highly correlated (Pearson's $R > 0.75$) between discovery and replication cohorts.

(A) Genes with the largest contribution to ambient transcript gene signature.



(B) Predictability of clusters at various resolutions.

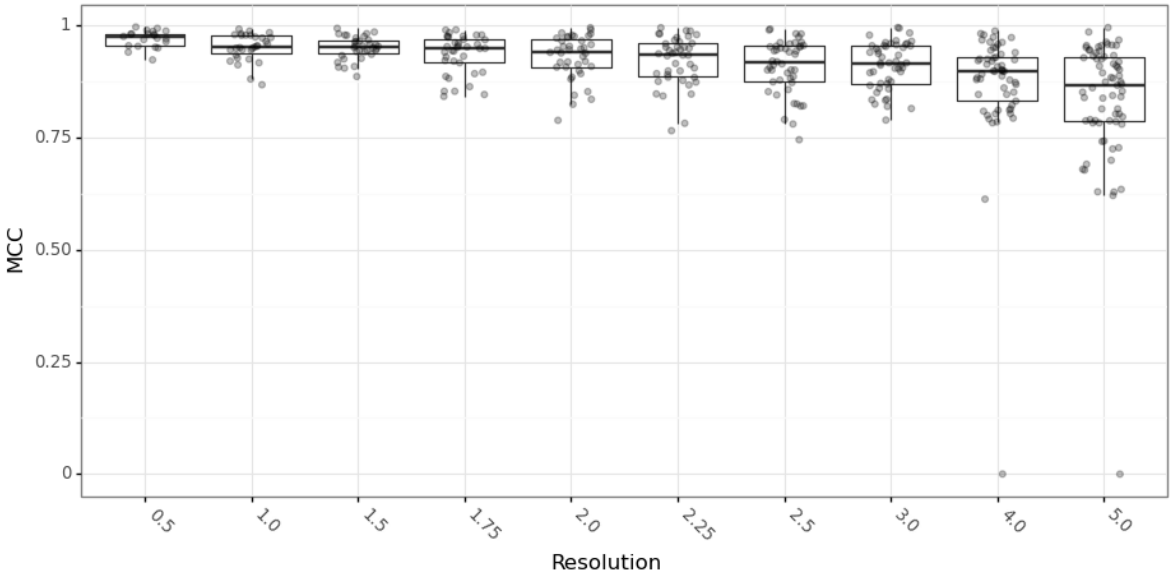


Figure S9

Figure S9. Quality control and cluster optimisation.
(A) The 75 genes (x axis) identified by CellBender that contributed most to the ambient transcript signature (y axis) across samples in the discovery dataset. (B) Cluster resolution (x axis) and the predictability of clusters (points) using cells withheld from training as measured by Matthews correlation coefficient (MCC; y axis; Methods). We selected a resolution of 3 as the predictability of clusters rapidly deteriorates at resolutions $>$ 3.