Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Spatial constraint and transient reversion initiate intestinal regeneration

Douglas Winton

doug.winton@cruk.cam.ac.uk

Cancer Research UK Cambridge Institute https://orcid.org/0000-0001-6067-7927

Doran Khamis

MRC Weatherall Institute of Molecular Medicine

Atanasiu Demian

MRC Weatherall Institute of Molecular Medicine

Elisa Moutin

University of Cambridge CRUK Institute https://orcid.org/0000-0002-9029-1520

Giada Giavara

CRUK Cambridge Institute

Mathilde Colombé

CRUK Cambridge Institute

Sarah Kozar

CRUK Cambridge Institute

Richard Kemp

Cancer Research UK Cambridge Institute

Michael Morgan

Institute of Medical Sciences https://orcid.org/0000-0003-0757-0711

Edward Morrissey

MRC Weatherall Institute of Molecular Medicine

Article

Keywords:

Posted Date: August 27th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10663694/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: There is **NO** Competing Interest.

Spatial constraint and transient reversion initiate intestinal regeneration

Doran Khamis*[2], Atanasiu Demian*[2], Elisa B. Moutin*[1], Giada Giavara[1], Mathilde Colombé[1], Sarah Kozar[1], Richard Kemp[1], Douglas J. Winton[1] ✉, Michael D. Morgan[3]✉, Edward Morrissey[2]✉

[1]CRUK Cambridge Institute, University of Cambridge, Cambridge, CB2 0RE, UK
[2]MRC Weatherall Institute of Molecular Medicine, University of Oxford, John Radcliffe Hospital, Headington, Oxford OX3 9DS, UK
[3]Institute of Medical Sciences, School of Medicine, Medical Sciences and Nutrition, University of Aberdeen, Aberdeen, AB24 3FX

*These authors have contributed equally

Correspondence
✉ doug.winton@cruk.cam.ac.uk
✉ michael.morgan@abdn.ac.uk
✉ edward.morrissey1@astrazeneca.com

ABSTRACT

The cellular origins of intestinal stem cell regeneration in the crypt base after localised damage is uncertain due to an incomplete quantitative understanding of the interplay of the cell behaviours that maintain and restore intestinal crypts. Here, machine learning processing of combined lineage tracing and comprehensive profiling of single cell transcriptional data assigns crypt position and cell cycle status to each intestinal crypt cell. This allows population dynamics models to be applied to provide quantitative insights into short- and long-term cellular behaviours in homeostasis and after intestinal stem cell ablation. This reveals that passive retrograde movement previously implicated in the homeostatic maintenance of the intestinal stem cell pool is not sufficient for its restoration after ablation. Rather, 10% of the regenerated stem cell pool is seeded from a localised suprabasal cell population marked by acquisition of foetal gene expression, including *Ly6a*, with 90% of stem cells being restored subsequently by expansion of this subpopulation.

INTRODUCTION

The simple columnar epithelium of the small intestine differentiates along a linear axis, starting at the base of glandular crypts and ending at the tips of projecting villi. Along this axis, functionally defined zones comprising stem, proliferative and terminal differentiated compartments are maintained at a constant size despite high rates of cell proliferation and rapid vertical mass migration[1]. Absorptive and secretory progenitors created from intestinal stem cells (ISCs) first arise near the crypt base, before differentiating into either mature enterocytes, goblet, enteroendocrine or tuft cells that are vertically displaced. Maturing secretory Paneth cells are the exception, as they migrate downwards and contribute to the ISC-maintaining niche[2]. Within the stem cell zone, there is ongoing loss and replacement of ISCs [3,4], to which cells at the upper boundary can contribute due to downward or retrograde movement[5].

When steady state ISCs, typically *Lgr5*-positive, are destroyed following either radiation treatment or selective ablation, surviving cells act to restore them[6]. Controversy exists about the primacy of the identity and state of the surviving cells in mediating regeneration. Both quiescent cells, that principally locate immediately supra-basally to the stem cell zone (cell position “+4”), and a more general dedifferentiation of mature cells have been implicated[7–14].

The limitations of existing experimental methods have been highlighted, such as the use of a limited set of cell identity markers and a general acceptance, without challenge, of the existing ISC model in which most of the stemness potential lies in *Lgr5*+ Crypt Base Columnar cells (CBCs), with minor contribution from populations outside of the crypt base restricted to the context of regeneration after CBC loss[15,16].

This has led to an alternative model based on computational analysis of single cell transcriptomic data and chromatin accessibility[15,16]. These studies propose that a subset of rapidly cycling transit-amplifying cells (renamed isthmus cells or *Fgfbp*1+ upper crypt cells) contribute to the ISC pool in the steady state and mediate regeneration after ISC ablation. They conclude that most of the stemness potential is found outside of the crypt base, and that crypt restoration after depletion of Lgr5+ ISCs is a passive process. However, a more recent study employing similar approaches

and using mathematical modelling to integrate cellular behaviours in homeostatic crypt renewal reached the opposing conclusion, providing evidence for stem cells at the crypt base as the main driver of intestinal epithelium homeostatic renewal[17].

To reconcile conflicting observations and stem cell models, we combine single-cell transcriptomics profiling and lineage tracing to develop CRISP (Crypt Reconstruction Integrating Spatial and Proliferative states), a machine learning model that integrates single cell transcriptional data from absorptive and secretory lineages to create a unified, gland-level view of intestinal crypt maintenance in homeostasis and restoration after ISC ablation.

CRISP outputs were used in a population dynamics model providing quantitative insights into long term cellular behaviours in the intestinal crypt. This revealed that only limited retrograde movement in homeostasis is consistent with mass vertical migration. The implication is that retrograde movement from a population of cells outside of the crypt base cannot sustain epithelial self-renewal. The rate of retrograde movement is increased up to 10-fold during regeneration, but we find that only a subpopulation of cells immediately adjacent to the crypt base have a substantial probability of reversing their trajectory to replace ISCs.

Leveraging single-cell gene expression profiling post-ISC ablation identifies that tissue remodelling is characterised by gene expression changes associated with foetal development, and a specific cell cycle state previously linked with replication licensing. These transcriptional changes converge on a *Ly6a*$^{+}$ population within the subrabasal crypt region, providing a prime candidate cell state that drives ISC regeneration post-ablation.

RESULTS

Constructing a homeostatic single-cell atlas of the intestinal crypt

The cellular activities allowing non-ISCs to contribute to the ISC pool or to restore it after ablation are here termed homeostatic and regenerative reversion, respectively. To study both processes, murine single-cell RNA-sequencing (scRNA-seq) data were

used to transcriptionally profile homeostatic epithelial maintenance as well as changes in cellular responses at several timepoints following ISC ablation (Supp. Fig. 1A, B).

To establish a comprehensive atlas of intestinal crypt cell type heterogeneity in homeostasis, scRNA-seq was performed on crypt-enriched cells isolated from a panel of complementary transgenic mouse models and wildtype mice (Methods; Supp. Fig. 1A). FACS enriched populations included: total wild-type epithelial cells (EpCAM+), crypt-cell-enriched (EpCAM+/CD24+) and ISC-enriched (EpCAM+, *Lgr5*^{GFPDTR}+) populations. In addition, to capture and trace the major epithelial cell lineages in the small intestine, tdTomato^{+} and tdTomato^{-} cells were collected from *Atoh1*^{CreERT2}*; Rosa26*$^{TdTom(lsl)}$ mice[13] following tamoxifen induction. Both tdTomato^{+} and tdTomato^{-} cells, representing secretory and absorptive lineages respectively, were analysed between 24 and 72h after tamoxifen administration. Data measured under homeostatic conditions were collected into an integrated reference atlas and clustering performed. All expected cell types could be identified based on known markers expressed in the different clusters, including secretory (*Muc2*), absorptive (*Alpi*), Paneth (*Lyz1*), and ISCs (*Lgr5*) (Supp. 1C, D).

Spatial deconvoluton of cell cycle and lineage status from integrated single-cell transcriptomes

To infer dynamic cell movements within the crypt, information on cell cycle state and cell position in homeostasis and regeneration was required. A common spatial axis containing estimates of cell cycle status was derived from a diverse atlas of the crypt base epithelium using a multi-stage machine learning (ML) approach (Fig. 1A). First, the cell cycle status for all cells was derived using a custom ML model (termed ccVI; see Methods) that fits gene expression changes for a curated training set of 207 cell cycle genes that were compiled previously[18,19] (Supp. Table 1). This allowed every cell to be assigned a continuous cell cycle progression value. The second machine learning stage takes as input the expression of a cell, its cell cycle state, ablation or homeostatic experimental conditions, time point post-ablation, batch, and technology. The model, termed CRISP (Crypt Reconstruction Integrating Spatial and Proliferative states), maps the cell to a trajectory in latent space that is common to all experimental conditions. Once fitted to the whole data set, the spatial axis is binned into discrete

positions to infer a location along the crypt axis. Cell location and cell cycle status then allows fitting of rates of cell production and movement of cell mass through the crypt in both steady state and ablation conditions over time to provide a description of population dynamics.

Cell positions were assigned by the CRISP model using the expression of all genes and along both absorptive and secretory lineages (Fig. 1B). Of note, as the primary goal is to resolve the contribution of major intestinal epithelial lineages to the ISC pool, enteroendocrine and Tuft cell populations were disregarded. To ensure a robust crypt axis with approximately equal sampling of cells within assigned crypt regions, cell positions were binned to crypt base (CB; cell positions 1-4), lower crypt cells (LCC; cell positions 5-7), mid-crypt cells (MCC; cell positions 8-11 and upper crypt cells (UCC; cell positions 12-14). The accuracy of the estimated crypt positions was assessed by considering genes with known patterns of differential expression within the crypt (Fig. 1C). The canonical ISC marker *Lgr5* shows appropriately restricted expression in positions 1-4 (Fig. 1C, D). Recently proposed markers for the isthmus population, *Fgfbp1* and *Atad2*, appear more highly expressed in positions in the lower and mid-crypt regions than in the crypt base (Fig. 1C, D).

Genes with predicted spatially restricted patterns of expression by the CRISP model include *Mptx2*, Sox4, *Dmbt1,* and *Krt20*, with high expression in positions 1-4, 5-7, 7+ and 15+, respectively (Fig. 1E). RNAscope spatial mRNA profiling validated the expression pattern of these genes along the crypt axis both in homeostasis and post-ISC ablation (Fig. 1F), confirming the accuracy of the cell positional assignment from single-cell data.

Assessment of CRISP cell cycle assignments in homeostasis

In performing spatial assignments, CRISP minimises the impact of cell cycle effects. However, the model retains cell cycle information allowing the proliferative status of cells to be explored in different crypt regions. A cell cycle progression estimate (CCPE) was computed for each cell by training the model with the curated set of 207 cell cycle genes mentioned above (Supp. Fig. 1E, see Methods), enabling the interrogation of cell cycle distribution across the crypt axis. After removing the cell cycle arrested

Paneth cells from this analysis, most crypt base cells were classified as G1, agreeing with previous reports that ISCs have a longer cell cycle due to a protracted G1 phase[1,20] (Fig. 1G). In contrast, lower and mid-crypt cells were more evenly distributed across the cell cycle, in line with continuous cell cycling and short G1 phases that are a defining feature of transit-amplifying (TA) cells. Notably, many upper crypt cells were classified as G2/M, likely reflecting cells exiting the cell cycle after their last division at the upper end of the TA compartment. These aggregate positional effects agree with the known functional zonation in cell cycle states along the crypt axis. To corroborate the CRISP cell cycle assignments, they were compared to experimentally determined estimates of the fraction of cells in S-phase. The distribution obtained was in good agreement to that observed *in vivo* by incorporation of EdU (Supp. Fig. 1F).

Having computationally separated cell position from cell cycle, the relative contribution of each to fluctuations in gene expression through the cell cycle in different regions of the crypt can be evaluated. Initially, any cell cycle correlated genes that have not previously been described as such were sought by using the neural network to fit smoothed curves for each gene separated by position and lineage. Clustering analysis was performed after filtering for genes with a significant correlation with the cell cycle (see Methods). Gene expression trajectories identified 3 major trends representing peak expression in G0/G1, G1/S and G2/M (Fig. 1H, I). Next, the analysis was focussed on genes that had the same expression profile in at least four of five crypt regions: CB-ISC, LCC-Absorptive, LCC-Secretory, MCC-Secretory and MCC-Absorptive, which identified 328 cell cycle correlated genes (Supp. Table 2). Within this gene set, 142 overlapped with the 207 known cell cycle genes used to train the model (Fig 1H, I; Supp. Table 1). The large majority of the remaining 186 genes identified by this analysis had peak expression in G1/S (132 genes, 71%) with gene ontogeny (GO) analysis supporting this phase assignment (Supp. Fig1G; Supp. Table 3). Additionally, 31 genes assigned to G2/M appropriately generated GO enrichment for “nuclear division” and “organelle fission”. The remaining 23 genes assigned to G0/G1 phase were under-powered for enrichment analysis, however, several of these genes have been previously linked to cell cycle licensing (*Nrdg1*, *Sgk1*, *Lbh*)[21–23], histone demethylases linked to replication (*Kdm5b*, *Kdm7a*)[24] and growth factor signalling (*Pdgfa*, *Tcf4*)[25,26].

In a separate analysis, cell cycle regulated genes which showed differential expression during the cell cycle in ISC, LCC and MCC populations were compared, to assess if gene usage relates to different modes of division in TA-like compared to stem cell-like cell populations. The comparison showed 39 cycling genes unique to ISCs present in the crypt base, 19 unique to MCC and 7 unique to LCC (Fig. 1J). The large majority of cycling genes unique to ISCs appear as G0/G1 (30/39 genes, 77%). This set includes *Gkn3*, a secreted cell cycle inhibitor, as well as multiple Wnt and BMP signalling related genes (*i.e. Nr2e3, Shisa2*, *Lgr5, Rgmb)* likely regulating the known slow cell cycle progression of ISCs[1,20] (Supp. Table 3) (Fig. 1K). Of note, the previously described isthmus marker genes *Atad2* and *Stmn1* are primarily expressed in G1/S across all 3 crypt regions, suggesting that they may mark actively cycling cells rather than TA cells specifically.

Together, these analyses highlight the ability of CRISP to deconvolve spatial and cell cycle patterns, allowing genes linked to the cell cycle to be identified across diverse cell lineages and crypt positions.

Recreating crypt growth and migration dynamics with a dynamic population model

To evaluate the ability of CRISP to accuractly infer reversion dynamics, we first used scRNA-seq coupled to lineage tracing of *Atoh1* secretory-biased progenitors using tamoxifen-induced tdTomato (tdTom) expression in *Atoh1*^{CreERT2}*; Rosa26*$^{TdTom(lsl)}$ mice (Supp. Fig. 1A). As expected, the majority of *Atoh1* expressing single cells mapped to Paneth and secretory branches of the atlas UMAP (Fig. 2A). Single (24h) and multiple pulse (72h) protocols were designed to first label secretory progenitors and then capture cells that returned to an immature state and subsequently adopted a non-secretory fate. Applying CRISP across both sets of experiments, the proportion of ISCs expressing tdTom for each inferred crypt region was computed (Fig. 2B). This identified that total tdTom+ cells in the crypt increased from 0.4% at 24h to 1.8% at 72h, with a > 3 fold increase in tdTom+ cells in the crypt base (Fig. 2C). This analysis shows that CRISP captures the established activity of *Atoh1*+ cells and their contribution to homeostatic reversion[13,14].

To quantitatively characterise the dynamics of cell turnover and putative reversion, a mathematical population dynamics model was developed to estimate the rates of cellular transitions from the single-cell data (see Methods). First, to establish cell division rates, wild-type mice were injected with nucleotide analogues CldU and EdU with a two-hour separation to label cells in S-phase (Supp. Fig. 2A). This permitted an estimate of S-phase duration of around 6h and demonstrated that equivalent numbers of cells entered and left S-phase, shown by equal representation of $CldU^+/EdU^-$ and $CldU^-/EdU^+$ cells respectively, regardless of position (Supp. Fig2. B, C). Cell cycle times based on this estimate and the outputs of the CRISP model, including cell cycle progression estimates, were then obtained for all positions and cell types (Supp. Fig. 2D) (see Methods). Estimated cell cycle rates for ISCs (0.76 div/cell/day) were calculated to be lower than for LCC (1.47 div/cell/day), MCC (1.93 div/day) and UCC (1.28 div/cell/day), with the peak rate in MCC (Supp. Fig 2D) corresponding closely to directly determined published estimates of TA cell division rates[27].

Using the above cell division values, the population dynamics model was simultaneously fitted to both homeostatic 'snapshot' data and the time-related *Atoh1* lineage tracing data to estimate the forward and backward (reversion) transition rates between crypt regions (Fig. 2D). To confirm the accuracy of the model, its capacity to predict Paneth cell turnover was assessed by comparing it to published data derived from *in vivo* lineage tracing[28]. These prior experiments used saturation lineage tracing to label all epithelial cells except Paneth cells, allowing the inheritance of the label into the Paneth population to be measured over time. This comparison confirmed that the model closely recapitulates the trend and timescale of Paneth cell turnover (Supp. Fig. 2E). As well as confirming model accuracy, this analysis shows the model can predict long-term population dynamics (1000 hours) from single-cell data. Further, the model also predicted the known increase in cell migration rate at higher crypt cell positions (Suppl. Fig. 2F) and the production of around 220 cells per crypt per day (Fig 2E), with whole crypt epithelium being turned over (except Paneth cells) in around 2-3 days (Fig. 2F), in accordance with published observations[29]

Having confirmed that the model captures the known temporal and spatial dynamic behaviours of cells within the crypt, its output parameters were used to investigate homeostatic reversion. Results indicate that for a single crypt, on average one absorptive cell reverts every 16h and one secretory cell every 5 days (Fig. 2G). However, when adjusted for population size, both absorptive and secretory cells have an equal probability of reversion. A negligible number of Paneth cells revert per day, with zero included within the credible interval. As the model estimates forward and backward rates of movement for all crypt positions, the reversion ability of cells located in different positions along the vertical axis of the crypt was also determined, showing the reversion only occurs close to the crypt base, with the probability becoming vanishingly small beyond positions 9-10 (Fig. 2H).

Next, the model was used to predict the potential of homeostatic reversion rates to fully restore the crypt following a notional stem cell ablation event. After setting the ISC population to zero, the emergent time evolution suggested that crypt base repopulation could occur but would be both protracted, taking up to 10 days, generating only a small fraction of the ISC population (Fig. 2I) and lead to stable micro-crypt structures of around 20 cells (Fig. 2J). Together, the results from the population dynamics model indicate that single-cell expression data that are spatially mapped to crypt space and fitted to derive optimal solutions for cell production and movement can accurately predict long and short-term homeostatic maintenance of the epithelium. Secondly, the modelling analysis suggests that homeostatic dynamics alone are insufficient to fully restore ISCs after their ablation, suggesting additional mechanisms are involved.

Finally, the model was exploited to perform *in-silico* lineage tracing from the isthmus markers *Fgfbp1*, *Lgr4* and *Stmn1* in the steady state. Despite a broader expression pattern along the crypt axis than the ISC marker *Lgr5*, isthmus genes also labelled cells in the crypt base (Supp. Fig. 2G). This was not the case for the secretory cell marker *Atoh1*. As a result, when quantifying the ratio of labelled to unlabelled stem cells after 120h to that initially labelled, values were similar for the ISC marker *Lgr5* and isthmus genes, whilst this ratio was above 1 for *Atoh1* (Fig. 2K). This data suggests that markers intended to identify the "isthmus" population behave similarly to stem cell markers. Additionally, *in-silico* lineage tracing from the cell cycle gene

Mik67, largely expressed in TA cells, produces a similar ISC repopulation ratio as isthmus genes. This supports the idea that isthmus genes mark a population of highly-cycling cells above the crypt base.

Experimental ablation of CBCs

The crypt base can fully repopulate within 72h after injury[30]. To capture the regenerative response under conditions of minimal damage, ISCs were ablated using the diptheria toxin receptor (DTR) transgenic model, in which toxin receptor expression is driven by the *Lgr5* locus (*Lgr5*$^{GFP\ DTR}$). Diphtheria toxin (DT) was administered as a single dose, and single cells were isolated from *Lgr5*$^{GFP\ DTR}$ or *Lgr5*$^{GFP\ WT}$ animals after 12h, 18h, 24h, and 48h (Supp. Fig. 1B). Single-cell RNA-sequencing data from the ISC, absorptive and Goblet branches were processed with CRISP. Mapping cells onto the CRISP UMAP space confirmed that ISCs were depleted at 12h, and fully restored by 48h after DT treatment, as described previously[30] (Fig. 3A-C).

To identify genes responsive to ISC ablation along the crypt axis, differential gene expression analysis for each position and timepoint was performed (Methods, Supp. Table 4). 12h post-ISC ablation, differentially expressed genes (DEGs) in the villus indicated increased cell proliferation, with increased expression of *Ccnd2*, characteristic of G1/S transition (Fig. 3D). Up to 24h, a general shift was seen in all positions above the crypt base towards mucus production (*Agr2,* folding and trafficking of mucin, *Muc2*, mucin) and detoxification (*Gsta1* and *Gsta4*, glutathione transferase acting on electrophilic compounds). At 24h, fewer differentially expressed genes were detected, but were related to cytoskeletal remodelling for cellular transport, turnover and migration, denoted by increased expression of genes such as *Tubb4* and *Dynll1*. Reinforcement of the epithelial barrier was also seen in the crypt base with expression of *Cldn4*. After 24h, an anti-microbial response was mounted above the crypt base, characterised by expression of genes such as *Reg3g* and *Reg3b*. Interestingly, several genes part of a foetal signature previously linked with intestinal epithelial regeneration after injury were differentially expressed after ablation in LCCs (i.e. *Ly6a, Areg, Clu*)[31].

The effect of ISC ablation on the cell position mapping inferred by CRISP was tested by comparing expression of CRISP position markers at 12h and 24h post ISC ablation

(Fig. 3E), This showed that the cell positions inferred by the model hold true post ISC ablation. Next, analysis of the relative contribution of different crypt regions to the total number of differentially regulated genes post-ablation (Fig. 3F), identified that the highest number of transcriptional changes occurred in the lower crypt (positions 5-7). This pattern persisted until 24 hours, after which the total number of differentially genes had decreased by >5 fold and the highest number of DEG were identified in the newly restored CBCs.

Following recent reports proposing isthmus cells rather than *Lgr5*+ ISCs as the dominant progenitors of upper crypt cells and responsible for regeneration of the crypt base, expression of isthmus marker genes was probed at different timepoints post-ISC ablation (Fig. 3G). Of these *Fgfbp1*, *Stmn1*, *Lgr4* and *Birc5* showed no change in level or spatial distribution of expression when comparing steady state to 12h, 18h, 24h or 48h post-ablation. A slight increase in the expression of *Atad2* at 18h post-ablation was observed, which matched the increase in expression of the proliferation marker *Mki67*. This suggests that isthmus cells remain relatively passive in the process of ISC regeneration after ablation.

Gene co-expression identifies responsive sub-populations

The above analyses identified differentially expressed genes along the crypt-villus axis for the known epithelial lineages of the small intestine. However, this may miss the emergence of rare or transient populations of cells that are responsible for co-ordinating or driving regeneration, as previously postulated[7–14]. We reasoned that such cells would be identifiable by expression of distinct co-regulated gene modules. Hotspot is an open-source package that excels at finding heterogeneous cell states within larger cell populations[38]. 14 modules were identified through Hotspot in the ISC ablation scRNA-seq data, spanning different cell positions (Methods, Supp. Fig. 3A) and represented by enrichment in different functional annotation pathways (Supp. Fig. 3B). To identify subpopulations of cells likely to play a role in the regenerative response observed after ISC ablation, modules were ranked by the proportion of module genes differentially expressed 12 hours after ISC-ablation. The highest number of differentially expressed genes was seen in modules 12 and module 2 (Fig. 4A).

Next, the analysis focused on foetal genes, which were found to be expressed in lower crypt positions following ablation. Over-representation analysis was performed using the foetal signature previously identified by Mustata and colleagues[32]. This revealed significant foetal enrichment in module 12 compared to all other modules (Fig. 4B) (see Methods). Accordingly, single cell transcriptomic data showed enrichment in *Ly6a* and *Areg* expression in the lower crypt, overlapping with module 12 (Fig. 4C). To identify the contribution of module 12 cells to the increase in expression of foetal genes post-ISC ablation, the normalised expression of *Areg*, *Clu*, *Ly6a* and *S100a6* in module 12 cells was compared to the expression in all other cells (Fig. 4D). This revealed that the large majority of the increased expression in foetal genes post ISC-ablation is driven by module 12 cells. RNAscope probes showed a highly restricted upregulation of Ly6a-positive cells directly above the crypt base post-ISC ablation, which were not present in non-ablated controls (Fig. 4E). Notably the appearance of module 12 cells during regeneration suggests active adaption and contrast with the reported passive behaviour of the bulk of isthmus cell that were proposed to be inherently competent for long term self renewal.

Taken together, we find that there is a clear and well-located regenerative set of cells, located directly above the crypt base which is marked by expression of *Ly6a*.

G1-state in the lower crypt suggests licensing is involved in ISC regeneration

To determine how gene usage changes during regeneration, a similar analysis as before was performed, by fitting a smoothed curve for all genes against the CCPE to classify cells as significantly correlating with the cell cycle or not. First, the genes identified as core cycling genes in the homoeostatic analysis were considered (Fig. 1G-K) to reveal that of the 328 core cell cycle genes, the majority show consistent cycling patterns, with only 14 not appearing in all time points and positions (Supp. Table 5, Supp. Fig. 4A). This shows that this key set of genes are not only common to all lineages and positions in homeostasis but are also prevalent and cycling-associated in regeneration.

Next, to identify genes that change cycling pattern during regeneration, genes correlating with cell cycle progression estimates during regeneration and that separate

ISC, LCC and MCC populations were identified and compared to their homeostatic reference (Fig. 4F, Supp. Fig 4B). Gene ontogeny analysis by crypt position identifies cell death processes for genes that gain cell cycle behaviour in ISCs following ablation, and developmental processes for genes that gain and lose this behaviour in LCCs (Supp. Fig 5). This analysis also shows that the majority of genes changing status are in G0/G1 cells. Further, while most cells across the whole crypt are assigned to S phase following ISC-ablation, a large proportion of cells in module 12 are assigned to G1 (30% at 12h post ISC ablation, 65% at 18h post-ablation) (Fig. 4G). Together, these observation suggest that module 12 cells may be undergoing replication licensing in preparation for cell division.

Population dynamics model identifies main source of regeneration

To determine where cell mass is produced within the crypt after ISC ablation, cycling rates and proportions of cells in S-phase were determined for different cellular positions at different times post ISC ablation, as previously done in homeostastic setting (Fig. 5A). This revealed an increase in cell cycle rates in LCC and MCC from 12h to 24h post ISC ablation, with no change in UCC (Fig. 5B). By 48h, all cycling dynamics have returned to homeostatic values. As expected, no ISCs could be observed at 12h, and the few observed at 18h were cycling at close to homeostatic rates. By 24h, these cells were cycling at the same high rate as the LCCs, returning to homeostatic values by 48h.

The cycling rates allow an estimate of excess cell mass that is produced relative to homeostasis. By 24h post ISC ablation, enough excess cells can be produced to overpopulate the crypt base (Fig. 5C). However, it is unclear how these dynamics relate to restoration of ISCs, as the observed ISC population in the 12-24h period following ablation is substantially lower than in the steady state (Fig. 5D).

Directly integrating cell position into reversion (Fig. 2G), the population dynamics model was adjusted so that all excess divisions produced by LCCs are directed to the base. Under this condition, ISC repopulation is still generated more rapidly than is observed (Fig. 5E). Another passive means of repopulating the crypt base would include retaining the homeostatic dynamics but rescaling the proportion predicted from

the homeostatic inference to account for the proliferative burst accompanying regeneration. In this scenario, the model indicates that higher cycling rates prevent crypt population collapse (Fig. 5E) and allow more substantial repopulation of the base compared to the homeostatic repopulation previously considered (Fig. 2E). However, the ISC population stabilises to around half of its homeostatic size, suggesting that there must be enhanced regenerative reversion beyond homeostatic reversion. These analyses highlight that the regenerative process is not simply passive, or dependent on the excess cell numbers created in the LCC/MCC, but rather an active process driven by a subset of cells.

Next, a scenario was considered in which ISCs are restored from the subset of the LCC population corresponding to the *Ly6a*+ cells forming module 12 described above. The number of these cells and their cycling rate was estimated from the single-cell data. Simulations show that ISCs would regenerate as observed experimentally, and stabilise to the appropriate number of cells per crypt (Fig. 5E). The other modules identified by Hotspot analysis were similarly fitted by the population dynamics model and confirmed that module 12 provides a substantially better fit than any of the other modules (Fig. 5F).

Origins of LCC reversion and ISC proliferation in regeneration

To explore the origin of the the fully regenerated ISC population from all available crypt cells, the proportion of cells reverting compared to the proportion subsequently generated by cell division was investigated. Analysis of the rate of regeneration over time split by origin shows that while reversion dominates initially, ISC division rate quickly surpasses it by around 18h (Fig. 5G, H). The population dynamics model predicts that the majority of the initial reversion comes from modules 12 (Fig 5I, J). Reversion accounts for around 10% of regenerated cells. This reversion is primarily driven by module 12 cells absorptive cells, which in turn replicate and repopulate the base (Fig. 5K, L).

Together, these findings show that despite the large wave of division in response to damage, most cells do not contribute to the regeneration of the crypt base. Rather, regeneration is seeded by a small and slow cycling subset of *Ly6*a+ cells located

directly above the crypt base, which produces around 10% of the homeostatic complement of ISCs that fully restore the stem cell population.

DISCUSSION

The contribution of cellular plasticity to intestinal regeneration has conventionally been assessed by clonal lineage tracing that marks individual non-stem cells of a given cell type or state, to demonstrate their capacity to give rise to long-lived clones comprising all intestinal epithelial cell types. Common to these tracing experiments is a low, but appreciable, baseline contribution of non-stem cells, suggesting that plasticity is a common feature of cells within the crypt[7–16]. This interpretation is supported by the similarity of chromatin profiles of enhancers and promoters in absorptive and secretory lineages along the crypt axis and implies a low barrier to dedifferentiation and common mechanism to confer a regenerative state[12,33]. Collectively, these features would also be consistent with a passive process and direct exchange between isthmus and stem cell states.

To date, the relative contribution of different intestinal cell types and states to retrograde movement has been hampered by several factors. First, most cell types are considered independently. Second, there is considerable variability in the efficiency of labelling different cells using different lineage tracing tools. Third, the contribution of long-lived clonal progeny to a lineage or pool of cells is measured qualitatively or semi-quantitatively, without considering the total size of the population of interest. Consequently, within steady state and regenerative intestinal crypts it is unclear which cells show genuine plasticity, with what frequency and efficiency, and from what distance from the stem cell population. Additionally, whether the process involves passive processes or modifying adaptive behaviours is not known. For these reasons, a quantitative and comprehensive population dynamic model of crypt maintenance and regeneration was required.

Using neural network integration of single-cell transcriptomic data in homeostasis, lineage tracing from the secretory progenitor marker *Atoh1*, and *Lgr5*+ ISCs ablation with diphtheria toxin, we created a spatial and temporally defined cell state map of the intestinal epithelium. Integrating this map with a population dynamics model

529 determined the relative contribution of different cell populations to the ISC pool in
530 homeostasis and after ISC-depletion. The model confirms that retrograde movement
531 in homeostasis is consistent with the bulk upward migration that acts to maintain the
532 tissue, allowing limited contribution to and maintenance of the ISC pool by cells directly
533 above the crypt base. The *Lgr5*+ ISC population is primarily responsible for self-
534 renewal, as has recently been shown by others[17]. Of note, each of the isthmus/upper-
535 crypt marker genes is also expressed in ISCs located at the base of intestinal crypts,
536 suggesting that the assignment of self-renewing activity by the isthmus population in
537 previous studies may not fully discriminate between isthmus/upper crypt and *Lgr5*+
538 ISCs.
539
540 The retrograde movement observed in homeostasis is insufficient to explain ISC
541 restoration after ablation. Rather, we identify an adaptive cell response spatially
542 restricted within the suprabasal zone which are characterised by altered cell cycle
543 characteristics and expression of a foetal transcriptional signature. The high proportion
544 of these cells in a G1-state suggests they may be licensing for replication[1,20]. Further,
545 the expression of a foetal transcriptomic program aligns with many accounts of foetal
546 reversion during intestinal regeneration, including slow-cycling revival stem cells
547 activated by YAP1 signalling[34]. Foetal reversion has been associated with response
548 and tolerance to stress, including from inflammation [35–37]. Dynamic modelling of this
549 putative *Ly6a*+ renerative population indicates that their retrograde movement is
550 sufficient to initially seed *de novo* ISCs that subsequently proliferate to repopulate the
551 intestinal epithelium. The restricted size and specific transcriptomic signature of this
552 cell population suggest an adaptative, rather than passive regenerative response.
553
554 Together, our data supports a model where *Lgr5*+ ISCs located at the bottom of
555 intestinal crypts are the main contributors to cellular turnover in homeostasis.
556 However, when *Lgr5*+ ISCs are depleted, an adaptive regenerative response in a
557 subset of *Ly6a*+ cells located directly above the crypt base allows regeneration of the
558 intestinal epithelium. These findinds reposition cellular plasticity not as a passive
559 feature of isthmus cells, but as a latent program engaged in the case of an injury
560 ablating the intestinal stem cell pool. This provides a framework for understanding
561 regeneration across self-renewing epithelia.

ACKNOWLEDGMENTS

D.J.W. was supported by Cancer Research UK Cambridge Institute core funding grant A24456 and the research was primarily funded by the Wellcome Trust (grant 103805).

We thank the Histopathology Core, Genomics Core, Bioinformatics Core and the Biological Resources Unit at the CRUK Cambridge Institute for technical support and the members of the Winton, Morrissey and Morgan labs for their help.

AUTHOR CONTRIBUTIONS STATEMENT

D. K. processed the single cell transcriptomic data and performed the population dynamics modelling

T. D. processed the single cell transcriptomic data and performed the CRISP modelling

E. M. processed the single cell transcriptomic data and the RNAscope data, and wrote the manuscript

G. C. performed the single cell transcriptomic and CldU/EdU experiments

M. C. performed the single cell transcriptomic experiments

S. T. designed and performed the single cell transcriptomic experiments

R. K. processed the single cell transcriptomic data

D. J. W designed and interpreted the results of all experiments, and wrote the manuscript

M. D. M. designed and performed the single cell transcriptomic experiments, analysed and interpreted the results of all experiments and wrote the manuscript

E. M. designed, processed the single cell transcriptomic data and interpreted the results of all experiments and wrote the manuscript

COMPETING INTEREST STATEMENT

The authors declare no competing interests.

TABLES

Supplementary table 1: list of 207 cell cycle genes

Supplemetary table 2: consistent cell cycle genes

Supplemetary table 3: gene cell cycle status at different crypt positions

Supplemetary table 4: list of DEGs after ISC ablation and assignment to module

Supplementary table 5: gene cell cycle status at different crypt positions after ISC ablation

FIGURE LEGENDS

Figure 1: Crypt Reconstruction Integrating Spatial and Proliferative states. A) Schematic showing the structure of the Crypt Reconstruction Integrating Spatial and Proliferative states (CRISP) pipeline. B) UMAP representation of wildtype samples showing the expected cell types and assigned crypt regions. C) Dot plot showing the expression of lineage-defining genes and isthmus genes along the CRISP-assigned crypt axis. D) Average gene expression across crypt positions for isthmus genes shown. E) Average gene expression across crypt positions assigned by CRISP for genes shown that have known spatially restricted expression within the crypt. F) Image showing RNAscope detection of expression for spatially restricted genes shown at 12h and 24h post-ISC ablation in *Lgr5*^{GFPDTR} and wildtype control mice. G) Histograms of cell cycle progression rate (CCPE) in homoeostasis faceted by crypt region, highlighting a strong G1-skew in CBCs. H) Barplot showing numbers of genes with cell cycle correlated alterations in expression identified by CRISP. I) Output of cluster analysis showing cell cycle progression rate for genes with cell cycle correlated alterations in expression identified by CRISP (top = in training set, bottom = predicted). J) Number of genes showing differential expression during the cell cycle that are uniquely cell cycle regulated in the crypt population shown. K) Plot showing normalised expression across the cell cycle for a set of genes by crypt region.

Figure 2. Homeostatic population dynamics derived from single cell data. A) UMAPs showing cell types and representative expression of *Atoh1* in wild-type samples. B) UMAPs showing the location of tdTom+ cells 24h (Top)and 72h (top) after the start of tamoxifen induction. Box identifies the root (ISC) population shown in A. C) Plot showing the fold increase in proportion of *Atoh1*+ cells within each cell type and at each crypt region from 24h to 72h. D) Schematic representing the integration of modalities derived with either CRISP (single-cell data) or pulse chase with CldU and EdU into a population dynamics model that predicts rates of cell migration and reversion. E) Population dynamics model estimate of the number of cells created per

crypt per day, the red line indicates the expected number. F) In silico lineage tracing plot of the time taken for all crypt cells (excepting Paneth cells) to acquire label from ISCs due to cell turnover. G) Dot plots showing the number of cells reverting per day for each lineage in homeostasis (left) and the same values normalised by cell population size (right). H) Heatmap showing the probability of cells at different crypt positions reverting in the crypt base. I) Plot showing estimated ISC restoration by homeostatic dynamics after in silico ablation of stem cells. The blue line and shaded regions indicate number of ISCs generated with confidence intervals (excluding Paneth cells). J) Plot showing estimated number of cells in crypts predicted by homeostatic dynamics, after in silico ablation of ISCs. K) Ratio of labelled stem cells at T=120h post labelling compared to inital labelling for several isthmus genes, the ISC marker *Lgr5*, secretory marker *Atoh1* and cell cycle gene *Mik6.*

Figure 3: Single-cell spatial and proliferative mapping of regeneration after ISC ablation. A) UMAP showing cell types included in the ablation analysis. B) UMAPs mapping of the cells present at different timepoints post-ablation. C) Bar plot showing the proportion of cells per crypt region at different timepoints post-ablation. D) Bubble plot showing the maximum absolute log2 fold change (bubble size) of selected genes across crypt positions and timepoints post-ISC ablation. Only statistically significant changes are shown (p value < 0.05). E) Average expression of spatially restricted genes assigned by CRISP at 12h and 24h post-ablation, showing similar patterns compared to homeostasis. F) Barplot showing the proportion of differentially expressed genes at different timepoints post-ISC ablation and different crypt regions. G) Plot showing normalised gene expression depending on cell positions at different timepoints post-ISC ablation for the isthmus genes Atad2, Birc5, Fgfbp1, Lgr4 and Stmn1 as well as the proliferation marker Mki67. for Ly6a showing elevated expression in cells directly above the crypt base at 12h after ISC ablation.

Figure 4: Identification of responsive sub-population by gene co-expression. A) Bar plot showing the proportion of genes significantly differentially expressed at 12h after ablation in each module detected with Hotspot. B) Bar plot showing foetal gene enrichment across co-expression modules, ordered by significance. Bar height represents −log10(p-value) from a one-sided Fisher's exact test (foetal dataset from Mustata et al., 2013). C) UMAP projection showing enrichment for Hotspot module 12,

as well as expression of Ly6a and Areg at 12h post-ablation. D) Dot plot showing the normalised expression of foetal genes at different cell positions and timepoints post-ISC ablation in module 12 cells (blue) compared to all other cells (red). E) RNAscope RNAscope for Ly6a showing elevated expression in cells directly above the crypt base at 12h after ISC ablation. F) Number of genes showing differential expression during the cell cycle that are uniquely cell cycle regulated in the crypt population shown. G) Barplot showing proportions of Module 12 and LCCs (excluding Module 12 cells) in different stages of the cell cycle.

Figure 5: Population dynamics model pinpoints the source of ISC regeneration post-ablation. A) Heatmap showing the fraction of cells in S-phase at each cell position at different timepoints post-ISC ablation. B) Plot showing cycle rate for cells in different crypt regions at different timepoints post-ISC ablation. C) Bar plot showing the cumulative excess cell creation above that of the crypt base at steady state (green bar) at different timepoints post-ISC ablation (grey dotted line highlights the size of the ISC compartment at steady state). D) Frequency of ISCs observed following ablation. E) Line plot showing model prediction for the restoration of the ISCs population with time assuming it is mediated by: All excess divisions within the LCC population (solid line); the homeostatic rate of reversion rescaled by the excess cell divisions within the LCCs; Module 12 cells. Recorded stem cell proportions are shown with dots and error bars (red=steady state, blue=timepoints post-ISC ablation). F) Relative error in the population dynamics model fit when using each of the Hotspot modules as the regenerative population driving ISC restoration. G) Line plots showing model predictions for the rate of ISC restorations when originating from: Total reversion from the LCC population (blueline); Module 12 cell reversion; (purple line); homeostatic reversion rescaled by excess divisions in LCCs (green line); division of ISCs (red line). H) Line plots showing the cumulative proportion of cells reverting to the crypt base with modelling conditions as in G. I) Line plot showing model prediction for contribution of module 12 cells reversion to total restored ISCs at different timepoints post ablation. J) Stacked bar plot showing the number of cells from different branches reverting per hour at different timepoints post-ISC ablation. K,L) Bar plots showing number of crypts regenerated per 100,000 by: absorptive or secretory cells (K); by homeostatic reversion or the Module 12 cells (L).

REFERENCES

1. Schepers, A. G., Vries, R., Van Den Born, M., Van De Wetering, M. & Clevers, H. Lgr5 intestinal stem cells have high telomerase activity and randomly segregate their chromosomes. *EMBO Journal* **30**, 1104–1109 (2011).
2. Sato, T. *et al.* Paneth cells constitute the niche for Lgr5 stem cells in intestinal crypts. *Nature 2010 469:7330* **469**, 415–418 (2010).
3. Lopez-Garcia, C., Klein, A. M., Simons, B. D. & Winton, D. J. Intestinal stem cell replacement follows a pattern of neutral drift. *Science (1979).* **330**, 822–825 (2010).
4. Snippert, H. J. *et al.* Intestinal crypt homeostasis results from neutral competition between symmetrically dividing Lgr5 stem cells. *Cell* **143**, 134–144 (2010).
5. Azkanaz, M. *et al.* Retrograde movements determine effective stem cell numbers in the intestine. *Nature* **607**, 548–554 (2022).
6. Tian, H. *et al.* A reserve stem cell population in small intestine renders Lgr5-positive cells dispensable. *Nature* **478**, 255–259 (2011).
7. Tetteh, P. W. *et al.* Replacement of Lost Lgr5-Positive Stem Cells through Plasticity of Their Enterocyte-Lineage Daughters. *Cell Stem Cell* **18**, 203–213 (2016).
8. Jones, J. C. *et al.* Cellular Plasticity of Defa4Cre-Expressing Paneth Cells in Response to Notch Activation and Intestinal Injury. *Cell. Mol. Gastroenterol. Hepatol.* **7**, 533 (2018).
9. Yu, S. *et al.* Paneth Cell-Derived Lysozyme Defines the Composition of Mucolytic Microbiota and the Inflammatory Tone of the Intestine. *Immunity* **53**, 398-416.e8 (2020).
10. Yan, K. S. *et al.* Intestinal Enteroendocrine Lineage Cells Possess Homeostatic and Injury-Inducible Stem Cell Activity. *Cell Stem Cell* **21**, 78-90.e6 (2017).
11. Van Es, J. H. *et al.* Dll1+ secretory progenitor cells revert to stem cells upon crypt damage. *Nature Cell Biology 2012 14:10* **14**, 1099–1104 (2012).
12. Jadhav, U. *et al.* Dynamic Reorganization of Chromatin Accessibility Signatures during Dedifferentiation of Secretory Precursors into Lgr5+ Intestinal Stem Cells. *Cell Stem Cell* **21**, 65-77.e5 (2017).
13. Tomic, G. *et al.* Phospho-regulation of ATOH1 Is Required for Plasticity of Secretory Progenitors and Tissue Regeneration. *Cell Stem Cell* **23**, 436-443.e7 (2018).
14. Ishibashi, F. *et al.* Contribution of ATOH1 + Cells to the Homeostasis, Repair, and Tumorigenesis of the Colonic Epithelium. *Stem Cell Reports* **10**, 27–42 (2018).
15. Malagola, E. *et al.* Isthmus progenitor cells contribute to homeostatic cellular turnover and support regeneration following intestinal injury. *Cell* **187**, 3056-3071.e17 (2024).
16. Capdevila, C. *et al.* Time-resolved fate mapping identifies the intestinal upper crypt zone as an origin of Lgr5+ crypt base columnar cells. *Cell* **187**, 3039-3055.e14 (2024).
17. Banjac, I. *et al.* Fate mapping in mouse demonstrates early secretory differentiation directly from Lgr5+ intestinal stem cells. *Dev. Cell* **60**, 1281-1289.e6 (2025).

18. Whitfield, M. L. *et al.* Identification of Genes Periodically Expressed in the Human Cell Cycle and Their Expression in Tumors. *Mol. Biol. Cell* **13**, 1977–2000 (2002).
19. Kowalczyk, M. S. *et al.* Single-cell RNA-seq reveals changes in cell cycle and differentiation programs upon aging of hematopoietic stem cells. *Genome Res.* **25**, 1860–1872 (2015).
20. Carroll, T. D., Newton, I. P., Chen, Y., Blow, J. J. & Näthke, I. Lgr5+ intestinal stem cells reside in an unlicensed G1 phase. *Journal of Cell Biology* **217**, 1667–1685 (2018).
21. Amato, R. *et al.* Sgk1 activates MDM2-dependent p53 degradation and affects cell proliferation, survival, and differentiation. *J. Mol. Med.* **87**, 1221–1239 (2009).
22. Garikapati, K. *et al.* Blocking LBH expression causes replication stress and sensitizes triple-negative breast cancer cells to ATR inhibitor treatment. *Oncogene* **43**, 851–865 (2024).
23. Croessmann, S. *et al.* NDRG1 links p53 with proliferation-mediated centrosome homeostasis and genome stability. *Proc. Natl. Acad. Sci. U. S. A.* **112**, 11583–11588 (2015).
24. Gaillard, S. *et al.* Kdm5a and kdm5b histone-demethylases contribute to hu-induced replication stress response and tolerance. *Biol. Open* **10**, (2021).
25. Shoshkes-Carmel, M. *et al.* Subepithelial telocytes are an important source of Wnts that supports intestinal crypts. *Nature* **557**, 242–246 (2018).
26. Korinek, V. *et al. Depletion of Epithelial Stem-Cell Compartments in the Small Intestine of Mice Lacking Tcf-4*. http://genetics.nature.com (1998).
27. Cheng, H. & Leblond, C. P. Origin, differentiation and renewal of the four main epithelial cell types in the mouse small intestine I. Columnar cell. *American Journal of Anatomy* **141**, 461–479 (1974).
28. Ireland, H., Houghton, C., Howard, L. & Winton, D. J. Cellular inheritance of a Cre-activated reporter gene to determine Paneth cell longevity in the murine small intestine. *Developmental Dynamics* **233**, 1332–1336 (2005).
29. C.S. Potten. *Extreme Sensitivity of Some Intestinal Crypt Cells to X and 'Y Irradiation*. *Nature* vol. 269 (1977).
30. Metcalfe, C., Kljavin, N. M., Ybarra, R. & De Sauvage, F. J. Lgr5+ Stem Cells Are Indispensable for Radiation-Induced Intestinal Regeneration. *Cell Stem Cell* **14**, 149–159 (2014).
31. Yui, S. *et al.* YAP/TAZ-Dependent Reprogramming of Colonic Epithelium Links ECM Remodeling to Tissue Regeneration. *Cell Stem Cell* **22**, 35-49.e7 (2018).
32. Mustata, R. C. *et al.* Identification of Lgr5-Independent Spheroid-Generating Progenitors of the Mouse Fetal Intestinal Epithelium. *Cell Rep.* **5**, 421–432 (2013).
33. Fink, M. *et al.* Chromatin remodelling in damaged intestinal crypts orchestrates redundant TGFβ and Hippo signalling to drive regeneration. *Nat. Cell Biol.* **26**, 2084–2098 (2024).
34. Ayyaz, A. *et al.* Single-cell transcriptomes of the regenerating intestine reveal a revival stem cell. *Nature* https://doi.org/10.1038/s41586-019-1154-y (2019) doi:10.1038/s41586-019-1154-y.
35. Chen, L. *et al.* TGFB1 induces fetal reprogramming and enhances intestinal regeneration. *Cell Stem Cell* **30**, 1520-1537.e8 (2023).
36. Kirino, S. *et al.* Fetal reversion from diverse lineages sustains the intestinal stem cell pool and confers stress resilience. *Commun. Biol.*

https://doi.org/10.1038/s42003-026-09533-x (2026) doi:10.1038/s42003-026-09533-x.
37. Yui, S. *et al.* YAP/TAZ-Dependent Reprogramming of Colonic Epithelium Links ECM Remodeling to Tissue Regeneration. *Cell Stem Cell* **22**, 35-49.e7 (2018).
38. DeTomaso, D. & Yosef, N. Hotspot identifies informative gene modules across modalities of single-cell genomics. *Cell Syst.* **12**, 446-456.e9 (2021).

METHODS

Mice

Experiments were carried out on homozygous *Atoh1*Cre^{ERT2} or *Lgr5*^{GFPDTR} mouse lines of C57BL/6 background. For lineage tracing experiments, the mice were heterozygous for the reporter gene (*Rosa26*$^{tdTom/+}$). Additionally, the LSL-*Kras*^{G12D} (JAX008179) and *Dll1*Cre^{ERT2} alleles were used in some experiments. Genotyping was performed by Transnetyx (Cordova, TN, USA) using real-time PCR.

Animal husbandry

Animal care and procedures were performed at the Cancer Research UK Cambridge Institute Biological Resource Unit according to the UK Home Office under the authority of a Home Office project licence (PD5F099BE) approved by the Animal Welfare and Ethical Review Body at the CRUK Cambridge Institute, University of Cambridge. Mice were housed under controlled conditions (temperature (21 ± 2°C), humidity (55 ± 10%), 12h light/dark cycle) in a specific-pathogen-free (SPF) facility (tested according to the recommendations for health monitoring by the Federation of European Laboratory Animal Science Associations). Animals had unrestricted access to food and water. None of the mice had been involved in any procedure prior to the study

Single-cell library preparation and sequencing

Each well of a Chromium Single Cell B Chip (10X Genomics) was loaded with a total of 16,000 cells. Either the Chromium Single Cell 3’ v2, v3 or 3’/5’ Gel Beads-in-emulsion (GEM), Library and Gel Bead Kit and Chromium i7 Multiplex Kit (10X Genomics) was used for library preparation, according to the manufacturer’s instructions. In short, well number 1 on the Chromium Chip B was loaded with a mixture of the cell suspension with the GEM Retrotranscription Master Mix. Wells 2 and 3 were loaded with gel beads and partitioning oil, respectively. Nanoliter-scale GEMs containing the single cells to be analysed were subsequently generated using

the Chromium Controller (10x Genomics). Inside of each individual GEM occurs the simultaneous production of barcoded full-length cDNA from poly-adenylated mRNA. Silane magnetic beads were used for the recovery and clean-up of the pooled fractions following the fragmentation of the GEMs, after which the recovered DNA was amplified, and cDNA products were separated by size selection. To optimise the amplicon size for the generation of the libraries, the full-length cDNA generated from polyadenylated mRNA was enzymatically fragmented and subjected to size selection. Subsequently, addition of P5, P7, a sample index, and TruSeq Read 2 (read 2 primer sequence) via End Repair, A-tailing, Adaptor Ligation, and PCR were performed to construct the libraries prior to sequencing.

Capillary electrophoresis on a Tapestation (Agilent) was used to assess the quality of the libraries generated prior to sequencing. The libraries were sequenced using either an Illumina HiSeq 4000 or NovaSeq 6000 sequencing system (Illumina).

Single cell RNA-sequencing quality control and doublet removal

All scRNA-seq libraries were processed using Cellranger v4.10 using mouse genome build GRCm38. Non-empty droplets were called using emptyDrops88 (v1.12.3) on each multiplexed sample with 20,000 permutations, a lower UMI threshold of 100, and 1% FDR. Across all sequenced library samples, 113985 non-empty droplets were defined. We defined low-quality single cells as those with <1000 UMIs, and excessive mitochondrial RNA content as 5 median absolute deviations. Across all experiments we retained 98620 high-quality single cell transcriptomes. Doublets were identified using the scDblFinder (v1.1.8) package which generates artificial doublets by randomly sampling across the transcriptomes of observed single cells. Clusters with a high enrichment of artificial doublets were removed.

Crypt Reconstruction Integrating Spatial and Proliferative states (CRISP)

During the analysis of single cell data the cell cycle will often manifest as subclusters of the same cell type. When analysing different tissue conditions where the cell cycle is altered this effect can complicate integration and the interpretation of the data. The aim of the CRISP model is to integrate and align crypt cells from multiple experimental conditions, particularly for experiments where the cell cycle is altered. The model will first find the cell cycle state independent of cell type or experimental condition. In the

second step, cells are integrated across data sets into a single axis while minimising the impact of the cell cycle on this integration.

1. Cell cycle model

Cells undergoing cell division are observed to have oscillating gene expression throughout the cell cycle[17], with different key cell cycle genes peaking at different stages of the cell cycle. To capture this behaviour in our single cell data we model gene expression as a sequence of pulses. The pulses are implemented as linear transforms of cosines followed by a sigmoid. This combination produces pulses restricted to the (0-1) interval [fig s1a] which can then be used as a proportion to be used within a count based likelihood.

$$p = \sigma[A\,cos(\theta) + B]$$

The full pulse model has cell specific and gene specific parameters:

$$p_{cg} = \sigma[r_c r_g\,cos(\theta_c - \theta_g) + z_c z_g]$$

Where c indexes cells and g indexes genes. Here $\theta_g$is the point along the cell cycle axis θ where g peaks and θ_c the position along the cell cycle axis cell c has reached.

To fit this model to the single cell data we use a beta-binomial likelihood to directly model the raw single cell counts, with mean p_{cg} and dispersion coming from an overdispersed binomial.

$$x_{cg} \sim BetaBin(p_{cg}, N_c, \phi_g)$$

Where x_{cg} represents the counts for gene g in cell c, ϕ_g is a gene wise over-dispersion parameter and N_c corresponds to the read depth for cell c. We use a mean over-dispersion notation for the BetaBin.

To fit the model to the large number of cells available in the dataset we adopt a Variational Autoencoder set up, with a neural network as an encoder that receives the

counts for cell c and returns the cell specific variational parameters and as a decoder we use the pulse equation. Specifically the encoder is:

$$f_\theta(x_c) \rightarrow \mu_{r_c}, \sigma_{r_c};\ \mu_{\theta_c}, \sigma_{\theta_c}; \mu_{z_c}, \sigma_{z_c}$$

Where μ and σ are the variational mean and variance respectively.
The priors used for the cell parameters were:

$$r_c \sim N(1, 0.1), z_c \sim N(-5, 0.5), \theta_c \sim N(\pi, 4\pi)$$

All gene specific parameters were optimised during fitting.

The model was fit to a curated subset of 207 genes (see below for details of list) and mature cells were not used for fitting. fig s1b shows the fit for all cells, showing generally low posterior variance for θ_c, while fig s1c shows the fit for all genes, indicating where the gene is predicted to peak as well as its predicted peak value.

Genes used for fitting

We took two cell cycle gene sets from the literature, those of Whitefield et. al. (2002) and Kowalczyk et. al. (2015) [18,19], and combined their phase annotations. This initial long list of 795 genes was cut down to 203 genes by checking phase correlation scores in our single cell data. To this reduced list we added a small number of genes that were most enriched in the non-cycling subsets of the differentiating cells. The four genes Capns1, Cmpk1, Gfpt1 and Lgals4 showed the strongest enrichment, and so were added as “G0” genes to our cell cycle model gene list to act as indicators of a cell being out of cycle.

Analysing across genes and conditions for cyclical patterns

Once all the cells have an assigned cell cycle progression score we can assess the cycling behaviour of all genes and how that changes across lineage, crypt positions and conditions. To check for expression trends correlating with the cell cycle we fit separate smoothed curves to all conditions and genes. We do so by fitting the following model to the single cell count data

$$x^k{}_{cg} \sim BetaBin(p^k{}_{cg}, N_c, \phi_g)$$

$$p^k{}_{cg} = f_{kg}(\theta_{ck})$$

Here we have used a neural network f_{kg} to map the cell cycle progression value for cell c and condition k (crypt position, lineage and ablation time point) θ_{ck} to a vector with a proportion value for each gene $p^k{}_{cg}$. We additionally regularise these curves with a first and second order gradient penalty to smooth the curve and avoid fitting a trend unless there is a clear signal.

Next we filter the data to the genes and conditions for which the smoothed curve shows a significant trend. For all genes and conditions we evaluate $p^k{}_{cg}$ for a fixed number of points along θ. With this we calculate a coefficient of variation per gene-condition CV_{gk} and filter the genes per condition using Jenks Natural Break to partition the genes in the CV_{gk} distribution for a fixed k. We partition the distribution into 3, low CV (no cell cycle), intermediate and strong. We keep the intermediate and strong set and filter out the lowest.

Next we apply clustering to all the curves that remain after filtering. We consider each entity to cluster as a single curve gene-condition, meaning that the same gene will appear multiple times as different entities due to there being multiple conditions. We normalise all curves to the 0-1 interval and use a mixture model. The mixture model works by assuming that the data can be explained by a set of basis curves and that the gene curves are distributed with gaussian noise around them. We again use a neural network to create the curves, though here we do not regularise the curves. To capture many patterns we choose a large number of curves (15 curves) and regularise the coefficient of the mixture with a dirichlet prior as follows

$$\alpha_n \sim Dirichlet(\alpha_0 p_n)$$

$$p_n = \prod_n (1 - p_0)^n p_0$$

$$P(x_{gk}) = \sum_{n=1}^{N} \alpha_n \prod_i N(x_{gk}{}^i | \mu_{in}, \sigma_n)$$

$$\mu_{in} = g_n(\theta_i)$$

Here μ_{in} is value i for basis curve n calculated using neural network g. We normalise p_n to add to one. After fitting and inferring α_n and μ_{in}, we go back and assign each curve x_{gk} to a cluster. To simplify interpretation we collapse pairs of clusters where the curves barely change between them and when comparing two conditions many genes "flip" between these two clusters. Following this procedure leaves 3 clearly distinct clusters that match the pattern we would expect for G0/G1 accumulation, G1/S peak and G2/M peak.

2. Spatial integration model

<u>VAE Pseudospace Model</u>

Our framework builds up on the scvi model[39] which computes a representation of scRNA-seq data via Variational Autoencoders[40]. We sought to obtain a unified representation across multiple conditions (samples from ablation experiment, different sequencing technologies, batch effects etc.), therefore we implemented a conditional VAE model, applied on the gene expression dataset (*G* genes) that computes a low-dimensional latent representation (dimension *D*). This captures the biological signals independent of the effects from the various conditions. The encoder and decoder are both multi-layer perceptrons (MLP). We incorporate the cell specific information as covariates in the conditional VAE by appending the input to the encoder and decoder networks. For discrete covariates this takes the form of one-hot encoded vectors with size equal to the number of unique values; for continuous values we append one additional node.

We include the following covariates:

1. Experimental Batch (discrete)
2. Time after ablation - assumed continuous (0h, 12h, 18h, 24h, 48h).
3. Cell cycle component θ_c(continuous)
4. Sequencing technology (discrete)
5. Genotype (discrete)
6. Ablation status (discrete)

As the development of intestinal epithelial cells is a dynamic process with multiple cell states developing in parallel, we sought to extend the scvi implementation to incorporate these dynamics. We estimate the relative position of each cell in the developmental process, which corresponds to the physical position in the crypt-villus axis; we denote this term 'pseudospace'.

We compute the pseudospace trajectory in the latent space via neural ordinary differential equations[41]. We introduce a continuous function $S: R \rightarrow R^D$ in the latent space, whose input is pseudospace scalar s, and follows the following dynamics given by the differential equation:

$$\frac{dS}{ds} = h(S(s))$$

where the derivative function $h: R^D \rightarrow R^D$ is a multi-layer perceptron. The solution is computed via the solver developed by the authors (https://github.com/rtqichen/torchdiffeq/tree/master). Given pseudospace values $s_1, \ldots, s_N$ and initial state $S(s_1)$, the ODE solver returns $S(s_1), \ldots, S(s_N)$.

The epithelial cell development begins with a pool of ISCs that subsequently develop into separate cell populations. Therefore, we formulate a latent model that captures this behaviour, with the following assumptions. The cells in our dataset are split into 4 groups: ISC, Absorptive, Goblet and Paneth, and each group has their own trajectory in the latent space. The key assumption of our model is that Absorptive, Goblet and Paneth trajectories "branch off" from the end of the ISC trajectory, i.e. the last point of the ISC trajectory is the start point of the 3 trajectories. This enables us to obtain a continuous representation of the developmental trajectories. The conditional aspect of the VAE ensures we obtain a unified representation across the multiple conditions.

Let $i = 0,1,2,3$ denote the groups, with 0 indicating the ISC. Each group i has n_i cells with pseudospace values ${s_1}^{(i)}, \ldots, {s_{n_i}}^{(i)}$, sorted in increasing order, a corresponding trajectory function $S^{(i)}$ and derivative h_i. Each cell c has the corresponding covariate set $\boldsymbol{\mathcal{C}}_c$ as described above. We formulate the generative model as follows:

$${z_1}^{(0)} = S^{(0)}({s_1}^{(0)})$$

$$i \neq 0: {z_1}^{(i)} = S^{(0)}({s_{n_i}}^{(0)})$$

$$\forall\, i:\ S^{(i)}({s_1}^{(i)}),\ldots,S^{(i)}({s_{n_i}}^{(i)}) \;=\; ODESolve({z_1}^{(i)}, h_i,\ {s_1}^{(i)},\ldots,{s_{n_i}}^{(i)})$$

$$z_c{}^{(i)} \sim N(S^{(i)}({s_c}^{(i)}), diag({\tau_i}^2))$$

$$x_{cg}{}^{(i)} \sim NegativeBinomial(L_c\, f(z_c{}^{(i)}, \boldsymbol{\mathcal{C}}_c), \theta_g)$$

The latent variable ${z_c}^{(c)}$ of cell *c* has a generative distribution centered around the corresponding point on the trajectory $S^{(i)}({s_c}^{(i)})$, with variance parameterized by ${\tau_i}^2 \in {R_+}^D$.

Conditional decoder $f:\ R^{D+C} \rightarrow \triangle^{G-1}$ maps the latents to the unit *(G-1)*-simplex, such that the output represents normalized gene proportions. These are multiplied with the library size $L_c := \sum_g x_{cg}$ to obtain the predicted mean gene count. We formulate the mean-dispersion parameterization of the Negative Binomial distribution[39]. Gene-specific dispersion is quantified by parameter θ_g.

For the inference model, we assume a Normal variational posterior $q(z_c \mid x_c, \boldsymbol{\mathcal{C}}_c) = N(x_c;\ \mu_z, {\sigma_z}^2)$, where $\mu_z, {\sigma_z}^2$ is computed by the encoder $E_z:\ R^{G+C} \rightarrow R^{2\times D}$, which takes as input gene expression and covariates.

We also formulate $E_s:\ R^{G+C} \rightarrow [0,1]$ for computing pseudospace.The output is normalized to the unit interval. For cells from non-ISC groups we compute their pseudospace by adding ${s_{n_0}}^{(0)}$, i.e. pseudospace value of the last cell in the ISC trajectory, to the pseudospace encoder output.

The ELBO is given by

$$\boldsymbol{\mathcal{L}}(x, \boldsymbol{\mathcal{C}}) \;=\; \mathbb{E}[log\ p(x|z, \boldsymbol{\mathcal{C}})] \;-\; \beta\ KL(q(z|x, \boldsymbol{\mathcal{C}})\ ||\ p(z|S, \tau))$$

Scalar β controls the regularization between the variational posterior and generative distribution of the latents; higher values lead to a latent representation more tightly centered around the trajectory. We train the model using the Adam optimizer[39]. The resulting pseudospace values are a continuous approximation of the cell ordering on the crypt-villus axis. We estimate the discrete cell positions by distributing the sorted continuous values into 22 bins of equal cell numbers, with the reasoning that cells sorted in the experiments would be sampled uniformly at random from all positions.

Deriving the position markers

We sought to identify genes that predict pseudospace values well. Therefore, for each gene in the wild-type dataset we ran a random forest regressor between the pseudospace and expression values, and sorted genes by their coefficient of determination score.

Counterfactual differential expression

We restricted our analysis to the dataset from the ablation experiment, and additionally removed the Paneth cells. This contained expression data from multiple experimental batches, each consisting of control and post-ablation samples from at least one timepoint (12h, 18h, 24h, or 48h), a total of 5 batches. Two batches (17940 and 17994) have more than one post-ablation timepoint, while the remaining three consist of one timepoint each.

We sought to obtain a continuous representation of expression changes during genereration as a function of post-ablation time. A key ability of our VAE framework is to compute counterfactual predictions, i.e. predicting expression of a cell under different conditions. For example, predicting expression of a control cell had it been subjected to ablation, as a function of time and other covariates. We compute this by first mapping the observed expression into the latent space via the encoder, then mapping via the decoder with custom covariates back into the expression space. This allows us to obtain a continuous representation of regeneration by computing these counterfactual predictions for each cell across sequential timepoints, including control (homeostasis). As we have also estimated cell positions via our neural ODE model, we can capture both spatial and temporal regeneration effects.

As this representation is time- and batch- dependent, it would require each batch to have data from the control samples and multiple timepoints. To balance the batch effects, we adopted the following decoding scheme in our counterfactual framework:

(i) Batches 17940 and 17994 are decoded with their respective batch covariates

(ii) For the other 3 batches we compute their counterfactual predictions under the 17940 and 17994 covariates and then compute average expression between the 2 representations.

Once we computed this representation for all cells, we can identify gene expression changes specific to regeneration via differential expression (DE) analysis between control and post-ablation predictions across multiple timepoints.

3. Cell cycle phase partition

To assign individual cells to a cell cycle phase we use a mixture model on the inferred cell cycle progression estimates for all cells. The mixture is a Gaussian mixture model that has a different set of mixing coefficients for each cell position and lineage, while the mixing components are shared across all conditions and lineages. Specifically

$$P(x) = \prod_{c} \sum_{k=1}^{N} \alpha_{j,k} N(x_{c,j} | \mu_k, \sigma_k)$$

Where k is the index of the mixture model, c is the cell and j is the index that codes the lineage and cell position for cell c. For the fitting we use stan to estimate the parameters and use 10 mixing components. Once we have fit the mixture model we use the genes Pcna, Top2a, Ube2c, Mcm6 as markers to identify which mixture components belong to which phase. If normalised levels of Mcmc6 are high but Top2a is low we classify as G1. If the levels of Pcna and Top2a are high but Ube2c is low we classify as S phase. Finally if Ube2c is high and Mcm6 is low we assign as G2M. Finally, to identify non cycling components, if Mcm6 is low and Ube2c is low we classify as non-cycling. For global thresholds to separate cell cycle progression we aggregate all single cell data to derive phase value cutoffs.

Population dynamics model

Population model

The temporal evolution of the cell population in each position/lineage bin of the crypt, x, can be modelled using the dynamic system

$$\frac{dx}{dt} = A\,x,$$

where A is a matrix that contains the transition rates between the population bins. This equation has a general solution when A does not vary with time found by integrating with respect to time, t,

$$x(t) = x_0 exp\{A\,(t - t_0)\},$$

where $x_0$is the cell population distribution at some initial time $t = t_0$. In the situation where the transition matrix A was known and an initial cell population x_0 has been observed, this solution form can be used to predict the future dynamics of that cell

population. The aim of this modelling work, then, is to learn A from the observational data.

Steady state

In homeostasis the cell population does not vary with time and so $dx/dt = 0$. To learn A in homeostasis we must solve the steady state equation

$Ax = 0$.

Since we have observations of single cells from many different crypts, the absolute number of cells we have do not represent a single crypt population and so cannot be compared directly across experimental replicates. In order to model replicates together we can instead model the proportion of the crypt that is in each position/lineage bin,

$$c = x / \sum_{bins} x$$

where c sums to one by definition.

To infer c from the observed cell numbers x_i we use a BetaBinomial distribution for each replicate i,

$x_i \sim BetaBinom(N_i, c\,\phi, (1-c)\phi)$,

where N_i is the total number of cells observed in the ith replicate and ϕ is a precision parameter per bin that is shared across the replicates. Rather than treat the linear system $Ax = 0$ directly, which creates a very rigid sampling problem, we instead project our inferred homeostatic population forward in time and enforce a softer steady state condition through conservation of mass:

$$z_0 = N_0\, c$$

$$z = z_0 exp\{A\,\Delta t\}$$

$c \sim Normal(z/N_0, \sigma)$,

where σ is the strength of the mass conservation: how steady state the system is. A value of $\sigma = 1 \times 10^{-5}$ was used in the inference. The pseudo crypt population N_0 is just a normalization constant for the bin population z_0 and can be set to any positive number.

The missing piece of the puzzle is A. We start with the assumption that forwards movement (migration/differentiation) is the dominant dynamic in the crypt, with

backwards movement (reversion) being a rare event. With this in mind, we first treat the simplified system where no reversion occurs. Using the inferred branching lineage structure we can construct the master equation governing the transition of cells between the position/lineage bins which has the general structure

$$\Delta c_i = \gamma_i c_i - \lambda_{i->j} c_i + \lambda_{k->i} c_k,$$

where γ_i is the cell cycle rate of the ith bin (calculated using the cell cycle model described above) and $\lambda_{i->j}$ is the transition rate of cells from bin i to bin j. Since in steady state $\Delta c_i = 0$, we can rearrange to find the outward transition rate in terms of the cell cycle and the inward cell cycle rate,

$$\lambda_{i->j} = \gamma_i + \lambda_{k->i} \frac{c_k}{c_i}.$$

This system of transition rates can be reduced to only three free parameters governing the ratio of mass movement at the lineage branching point leaving the stem cells, which we can constrain as a three-simplex, S, with the prior

$$S \sim Dirichlet(1).$$

At sinks (the final cell position bin at the end of each lineage branch) mass leaves the system. The rate of transition out of a sink is defined by the same master equation rearrangement as above, but conservation of mass can be ensured by constraining one sink deterministically in terms of the others:

$$\lambda_{sink1->exit} = \sum_{i \in bins} \frac{\gamma_i c_i}{c_{sink1}} - \sum_{M \in sinks} \frac{\lambda_{sinkM->exit}\, c_{sinkM}}{c_{sink1}}.$$

Next, we introduce backwards movement, with soft shrinkage priors on the reversion rates, and infer the perturbation of the forward rates away from the forward-only transition rates:

$$\lambda^{(b)}{}_{j->i} \sim Beta(1, 10)$$

$$\lambda^{(\Delta f)}{}_{i->j} \sim Normal(0, \sigma_{\Delta f})$$

<u>Modelling lineage tracing in the continuum setting</u>

The Atoh+ lineage tracing experiments were key in helping to fit the parameters of the homeostatic model with backwards movement, which is underconstrained by definition. By observing the change in Atoh+ cells found in the absorptive lineage between the 24 hour and 72 hour timepoints, the scale of reversion of Atoh+ cells from

the secretory branch to the crypt base can be observed, thereby setting the overall magnitude of backwards movement in the homeostatic crypt.

Two pulse-chase procedures were used to collect Atoh+ lineage tracing data: (i) a single pulse with collection 24 hours later; and (ii) three pulses separated by 24 hours with collection 24 hours after the final pulse. To model the pulse labelling mechanism, a threshold κ was inferred that acts upon the (reconstructed) Atoh gene expression: if a cell has relative Atoh expression greater than κ then it is labelled by the pulse. A separate threshold κ_{pan} was inferred for the paneth cells to account for the apparent suppression of lowly expressed genes (like Atoh) due to the saturation of counts from highly expressed genes (like Lyz1). Beta priors were used for the thresholds κ and κ_{pan},

$$\kappa_i \sim Beta(\alpha_L, \beta_L),$$

where the parameters α_L and β_L, different for κ and κ_{pan}, were informed by the observed Atoh expression in the Atoh positive and Atoh negative cells from the lineage tracing experiment.

A labelling vector L, containing values in $[0,1]$ for each position/lineage bin, was constructed by thresholding the Atoh expression. Then, the experimental timeline was reconstructed as follows: the first pulse labelling is applied to the inferred homeostatic population x_{hon} to find the initial positive (labelled) and negative (unlabelled) cell populations:

$$x_{L1}{}^{+} = L\, x_{hon}$$

$$x_{L1}{}^{-} = x_{hom} - x_{L1}{}^{+}.$$

These populations are then simulated forward to the first observation time 24 hours later,

$$x^{+}(24) = x_{L1}{}^{+}\, exp\{24\, A\}$$

$$x^{-}(24) = x_{L1}{}^{-}\, exp\{24\, A\},$$

and the following two terms are added to the likelihood,

$$x_{obs}{}^{+/-} \sim BetaBinomial(N_{obs}{}^{+/-}, c^{+/-}(24)\, \phi^{+/-}, (1 - c^{+/-}(24))\phi^{+/-})$$

where

$$c^{+/-}(t) = \frac{x^{+/-}(t)}{\sum_{bins} x^{+/-}(t)}\ .$$

The second pulse labelling is calculated by using the intuition that the labelling can act on only the currently unlabelled portion of the crypt,

$$x_{L2}{}^{+} = x^{+}(24) + Lx^{-}(24)$$

$$x_{L2}{}^{-} = x^{-}(24) - Lx^{-}(24).$$

These populations are again simulated forward via

$$x^{+}(48) = x_{L2}{}^{+}\, exp\{24\, A\}$$

$$x^{-}(48) = x_{L2}{}^{-}\, exp\{24\, A\},$$

and this process is done a third time for the third label,

$$x_{L3}{}^{+} = x^{+}(48) + Lx^{-}(48)$$

$$x_{L3}{}^{-} = x^{-}(48) - Lx^{-}(48)$$

$$x^{+}(72) = x_{L3}{}^{+}\, exp\{24\, A\}$$

$$x^{-}(72) = x_{L3}{}^{-}\, exp\{24\, A\},$$

wherein the the following two terms are added to the likelihood to capture the observations from the three-day pulse chase experiment:

$$x_{3day\ obs}{}^{+/-} \sim BetaBinomial(N_{3day\ obs}{}^{+/-}, c^{+/-}(72)\, \phi^{+/-}, (1 - c^{+/-}(72))\phi^{+/-}).$$

Stochastic simulations of cell reversion

To calculate the probability of a cell starting in each position/lineage bin of reverting to the crypt base in homeostasis, we convert the continuous dynamic system described above into a continuous-time Markov process parameterised by position-dependent cell cycle rates and the posterior distributions over the forward and backward cell transition rates, λ and $\lambda^{(b)}$, described above.

Crypt compartments were represented as discrete states corresponding to cell positions and branches within the crypt architecture. An additional absorbing "Exit" state represented cell loss from the crypt. Transition events between the compartments were sampled using the Gillespie algorithm with exponential waiting times. At each simulation step, the propensity of every possible event was calculated as

$$a_i = n_i\, r_i,$$

where n_i is the number of cells eligible for the event and r_i is the corresponding transition rate. To stabilise the crypt size, a short refractory period was added to

division events so cells could not be chosen to divide until at least 30 minutes after a previous division by that same cell.

To initialise a simulation, a posterior draw was selected at random. The corresponding posterior cell abundance estimates were scaled to a total population of approximately 250 cells while preserving relative compartment frequencies, c, inferred in that same posterior sample. The number is approximately 250 as the number in each bin is rounded to the nearest integer. Individual cells were tracked explicitly as agents assigned to their current compartment.

The crypts are then simulated forward for four days, with the identity of cells entering the crypt base from above recorded. The probability of reversion is then calculated as the number of cells from each position/lineage bin that entered the crypt base in the four days divided by the number of cells from that position/lineage bin that were simulated. Results were obtained from 80,000 crypt simulations, representing approximately 2×10^7 cells.

Modelling lineage tracing in the discrete stochastic setting

Lineage-tracing simulations are performed using the same stochastic crypt model described in the previous section, with the addition of heritable cell labels. Initial labelling probabilities associated with lineage-tracing reporters are derived from the gene expression level outputs of the CRISP VAE pseudospace model described above.

Labelling probabilities for individual cells are derived from gene expression levels. Gene expression levels are first normalised independently for paneth and non-paneth populations using the 99th percentile of expression values. Normalised values below a threshold of 1/3 are set to zero and values above one were truncated to one. The resulting values L_{prob} are interpreted as probabilities that individual cells would be labelled at the start of the simulation.

To move from labelling probability for individual cells in the scRNA-seq data (*experimental* cells) to initial condition labelling in a ~250 cell crypt (*agent* cells), several steps are taken:

1. Set each *experimental* cell as labelled (i.e. value = 1, else 0 if unlabelled) if a sample $U(0,1) < L_{prob}$
2. Calculate the raw labelling proportion L_{frac} of each position/lineage bin by averaging the resulting label values
3. After creating the ~250 *agent* cell crypt with compartment frequencies as described in the previous section, generate the number of labelled *agent* cells by sampling a binomial for each cell type in each compartment $Binom(n_i, L_{frac})$.

Each cell carries its lineage label throughout the simulation. Migration events transfer the label with the cell, while division events duplicate the label into both daughter cells, thereby preserving clonal identity.

Crypt dynamics are simulated using the Gillespie algorithm as described above. Population states are recorded at regular six-hour intervals throughout the simulation by storing the full compartmental distribution of labelled and unlabelled cells. Simulations of 2,000 crypts are performed using independently sampled posterior parameter sets and random initial labelling configurations. The resulting ensemble of trajectories is used to characterise expected lineage-tracing dynamics while accounting for uncertainty in both crypt kinetics and initial reporter activation frequencies.

Validating model with orthogonal data

Paneth cell repopulation

Taking the fitted homeostatic population dynamics model, we ablated the paneth cells *in-silico* and simulated the dynamics of paneth repopulation. This stands as a good test of the ability of the model to accurately capture multiple timescales within a single framework, as it is known that the lifecycle of paneth cells is significantly longer than other crypt lineage. We compared our simulations to the paneth lineage tracing results from Ireland et. al (2005)[28] and found good agreement in the timescale of repopulation of the paneth niche.

Upwards migration rates

By calculating the net transition rates between cell position bins in the fitted homeostatic crypt model (forward - backward) and accounting for the binning of multiple cell positions together in the population dynamics model, predictions for the upwards migration rate of cells were made. The predicted migration rates, increasing from around 0.25 cell positions/hour at cell position 5 to a maximum of 0.9 cell positions/hour by cell position 15, agree well with the reported migration rates from Potten et. al. (1997) [29]. Our simulations also show the saturation of the migration rate above the point where cell cycling stops, capturing an intuitive result that was not predicted in Potten et. al. (1997) due to their observations not covering the upper crypt.

Number of cells created per day

By counting the number of cells leaving the top of the crypt in the fitted homeostatic model we can predict the productivity of the crypt in normal conditions. The commonly reported crypt productivity of 200 cells per day sits comfortably within the distribution predicted by the fitted model, showing our rendering of the cycling and migratory behaviours match expectations.

Time for turnover of whole crypt above base

By ablating *in silico* the entire crypt above the crypt base we can estimate the time it takes for the crypt base cells to replenish the crypt in homeostasis. Our simulations predict a turnover time of 3-4 days, again matching the commonly reported value.

Population dynamics in regeneration

To simulate crypt dynamics after ablation, we must move out of the homeostatic regime. However, the dynamics we learn from the homeostatic data can still form the backbone of our simulations, with three alterations:

1. Crypts exhibit drastic changes in cycling behaviour after suffering ablation.
2. The rate of cells migrating upwards and leaving the top of the crypt may change in step with the changes to the cell population and cell division within the crypt.
3. Backward movement (towards the base of the crypt) may be increased either locally or globally.

We aim to allow for these changes by making as few modelling assumptions as possible, and by deviating away from the homeostatic dynamics only as much as is required for regeneration.

To capture the changes in the cycling behaviour throughout regeneration we run the cell cycle inference model on the four ablation observation timepoints (12, 18, 24 and 48 hours post-ablation). In order to model the dynamics at a finer time resolution than the experimental timepoints, we then interpolate the inferred cell cycle rates between those calculated from the observational data, giving us a continuous estimate of the cycling rates at two-hourly timesteps.

To capture the changes in upwards cell migration and the rate of cells leaving the top of the crypt during regeneration, we calculate a transition rate multiplier based on the ratio to the homeostatic values of the current cumulative cell population and cumulative mass creation in the crypt. This creates a mechanism for the forward movement to be pushed away from the homeostatic values in response to the regenerative dynamics in the crypt. For example, at the start of the experimental timeseries the cumulative population of the crypt is decreased due to the ablation of the crypt base, meaning forward migration rates near the crypt base may be reduced to account for the lowering of mitotic pressure. In another scenario, if a simulation has fully repopulated the crypt base while the cell cycle throughout the crypt is still raised above homeostatic level, the forward rates adjust to allow the increased cell creation in the crypt to translate to increased productivity out of the top of the crypt, maintaining a realistic crypt size.

To capture potential increases in backwards movement we propose three modelling options: (i) backwards movement remains unchanged after ablation. This represents a passive regeneration, with homeostatic dynamics unchanged apart from the injection of extra mass through increased cycling. (ii) All excess cell division above the homeostatic rate is sent backwards. This represents a global concerted response to ablation while maintaining homeostatic dynamics as much as possible. (iii) All cell division from a subset of cells is sent backwards. This final option represents a targeted regenerative response in that it can only reasonably be followed as a strategy by a small subset of the crypt otherwise normal homeostatic functioning would break down. These three options suggest quite different regenerative behaviours. We use simulation experiments to ask which behaviour is most likely given the observational data.

The third option for retrograde movement is the one used for modelling regenerative modules. While the rest of the crypt undergoes "passive" regeneration through option (i), the cells identified as part of a putative regenerative module are given their own cell cycling behaviour distinct from the rest of the cells in their lineage/position bin, and the entirety of their cycling output is used for regeneration of the crypt base (i.e. translated into backwards movement).

Hotspot module analysis

We used Hotspot[38] to identify modules of correlated genes with regeneration specific effects. Therefore, we applied Hotspot on the filtered dataset containing post-ablation samples from timepoints 12h, 18h, 24h. We ran Hotspot using the latent representation from the VAE model as the input cell-cell similarity matrix, with 10 nearest neighbors and Negative Binomial null distribution. Hotspot computes local correlations on genes found to be autocorrelated (FDR<0.01), then computes modules via hierarchical clustering (we set min_gene_threshold=50, fdr_threshold=0.01, core_only=True). This results in 12 modules.

Hotspot computes assigns each cell a score, for each module. We select the cells enriched for module 12 by standardazing the module 12 score into a z-score (mean 0 and variance 1), then selecting cells for which z-score exceeded 3.

Module Enrichment Analysis

To assess whether each co-expression module was enriched for foetal intestinal genes, we performed an over-representation analysis (ORA) using Fisher's exact test. For each module, we defined the module gene set as all genes assigned to that module with available expression measurements. The universe of genes consisted of all genes detected in the dataset. Foetal intestinal genes were obtained from Mustata et al., 2013[32]. Gene symbols were standardized to ensure consistent formatting between datasets. For each module, we calculated the number of overlapping genes with the foetal gene set and constructed a 2×2 contingency table of overlap versus non-overlap relative to the universe. Fisher's exact test (one-sided, alternative = "greater") was used to compute the significance of enrichment, and odds ratios were reported to indicate effect size. Modules with no overlapping genes were assigned NA

for odds ratio and p-value. Significance was visualized as $-\log_{10}$(p-value), with the odds ratio indicated by color intensity in the plotted results.

RNAscope

Simultaneous detection of Mouse Mptx2, Krt20, Dmbt1, Sox4 and Ascl2 (panel 1), Top2a, Ly6a, Retreg and Gkn3 (panel 2) or Hist1h1e, Ly6a, Clu, Areg and Ascl2 (panel 3) was performed on FFPE sections using Advanced Cell Diagnostics (ACD) RNAscope® 2.5 LS Multiomic core reagent Kit (Cat No. 322930), RNAscope Multiomc C1, C2, C3, C4 and C5 Channel reagents (Cat No. 322935, 322940, 322945, 322950 and 322955, respectively),and RNAscope® 2.5 LS Probes (see table) (ACD, Hayward, CA, USA). Briefly, sections were cut at 3□M thick, baked for 1 hour at 60°C before loading onto a Bond RX instrument (Leica Biosystems). Slides were deparaffinized and rehydrated on board prior to pre-treatments using Epitope Retrieval Solution 2 (Cat No. AR9640, Leica Biosystems) at 95°C for 15 minutes, and ACD Enzyme from the Multiplex Reagent kit at 40°C for 15 minutes. Probes were visualized using Opal fluorophores (Opal 520, Opal 570, Opal 620, Opal 690 and Opal 780; Akoya Biosciences Cat No. FP1487001KT, FP1488001KT, FP1495001KT, FP1497001KT, and FP1501001KT, respectively) diluted using RNAscope LS Multiplex TSA Buffer (see table). Probe hybridisation, signal amplification and detection was performed on the Bond Rx according to the ACD protocol. Slides were then removed from the Bond Rx and mounted using Prolong Diamond (ThermoFisher Cat No P36965). The slides were imaged on the Akoya PhenoImager HT to create whole slide images. Images were captured at 40x magnification, with a resolution of 0.25 microns per pixel.

Target name-Channel	Cat No	Panel Number	Detecting Fluorophore	Fluorophore concentration
Mm Mptx2 - C1	467058	1	Opal 570	1:3000
Mm Krt20 - C2	402308-C2	1	Opal 520	1:1000
Mm Dmbt1 - C3	418568-C3	1	TSA-DIG & Opal 780	1:1000/1:250
Mm Sox4 - C4	471388-C4	1	Opal 690	1:1000
Mm Ascl2-C5	412218-C5	1	Opal 620	1:3000
Mm Top2a - C1	491228	2	TSA-DIG & Opal 780	1:1000/1:250
Mm Ly6a - C2	427578-C2	2	Opal 520	1:1000
Mm RetReg - C3	545858-C3	2	Opal 690	1:3000

Mm Gkn3 – C4	512068-C4	2	Opal 570	1:1000
Mm Hist1h1e – C1	584828	3	TSA-DIG & Opal 780	1:1000/1:250
Mm Ly6a – C2	427578-C2	3	Opal 520	1:1000
Mm Clu – C3	427898-C3	3	Opal 690	1:3000
Mm Areg – C4	430508-C4	3	Opal 570	1:1000
Mm Ascl2 – C5	412218-C5	3	Opal 620	1:3000

METHODS-ONLY REFENRENCES

39. Lopez, R et al. Deep generative modeling for single cell transcriptomics. *Nat. Methods.* **12**, 1053-1058 (2018)

40. Kingma, D. P. & Ba, J. Adam: A Method for Stochastic Optimization. http://arxiv.org/abs/1412.6980 (2017).

41. Cheng, J., Smyth, G. K. & Chen, Y. Unraveling the timeline of gene expression: A pseudotemporal trajectory analysis of single-cell RNA sequencing data. *F1000Res.* **12**, 684 (2023).

DATA AVAILABILITY

All single cell transcriptomics data used in this manuscript have been made publicly available on ArrayExpress under accession ID: E-MTAB-17022. Modelling and further analyses are publicly available on GitHub: https://github.com/ElEd2/paper-CRISP.

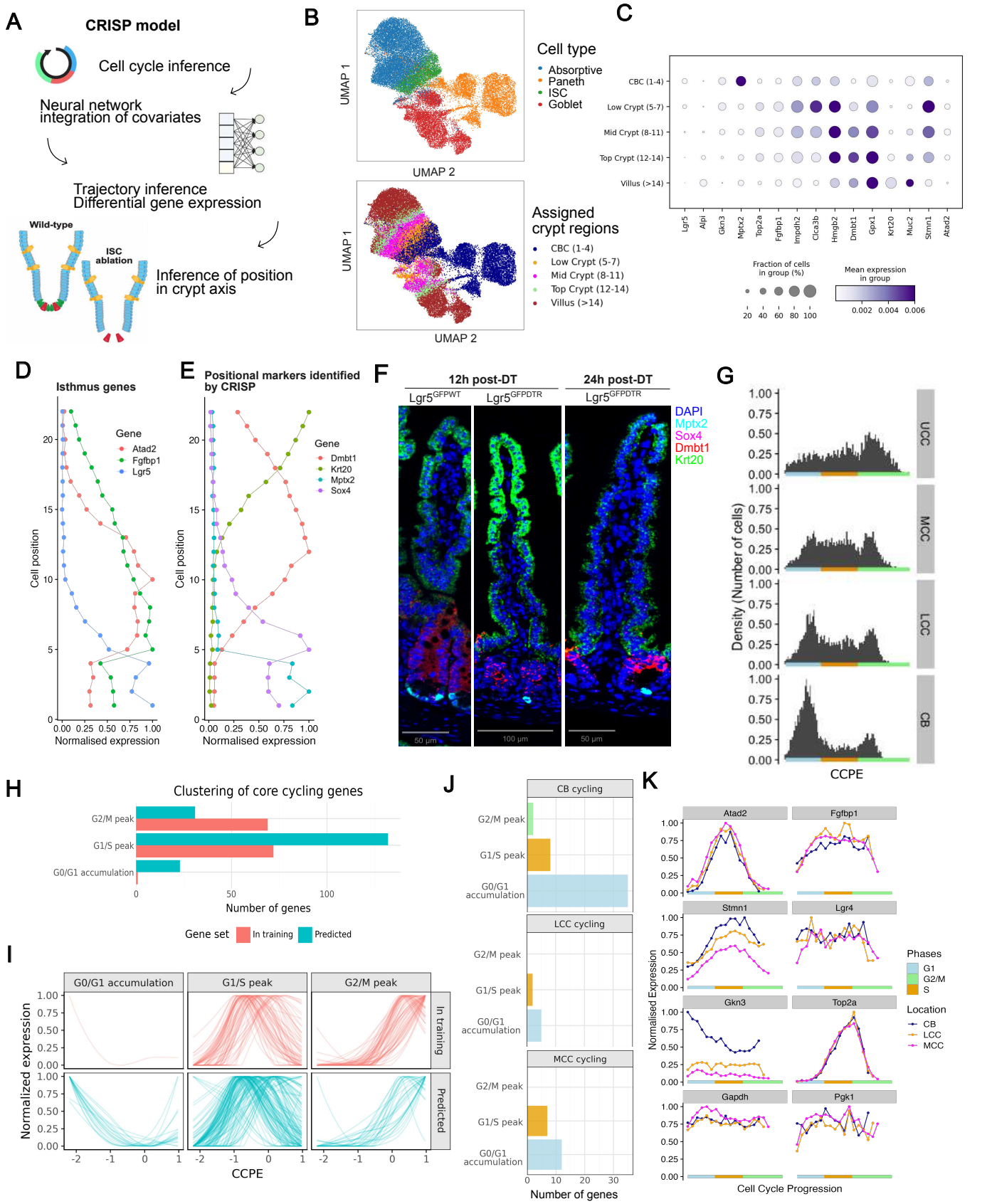


Figure 1: Crypt Reconstruction Integrating Spatial and Proliferative states. A) Schematic showing the structure of the Crypt Reconstruction Integrating Spatial and Proliferative states (CRISP) pipeline. B) UMAP representation of wildtype samples showing the expected cell types and assigned crypt regions. C) Dot plot showing the expression of lineage-defining genes and isthmus genes along the CRISP-assigned crypt axis. D) Average gene expression across crypt positions for isthmus genes shown. E) Average gene expression across crypt positions assigned by CRISP for genes shown that have known spatially restricted expression within the crypt. F) Image showing RNAscope detection of expression for spatially restricted genes shown at 12h and 24h post-ISC ablation in *Lgr5*^{GFPDTR} and wildtype control mice. G) Histograms of cell cycle progression rate (CCPE) in homoeostasis faceted by crypt region, highlighting a strong G1-skew in CBCs. H) Barplot showing numbers of genes with cell cycle correlated alterations in expression identified by CRISP. I) Output of cluster analysis showing cell cycle progression rate for genes with cell cycle correlated alterations in expression identified by CRISP (top = in training set, bottom = predicted). J) Number of genes showing differential expression during the cell cycle that are uniquely cell cycle regulated in the crypt population shown. K) Plot showing normalised expression across the cell cycle for a set of genes by crypt region.

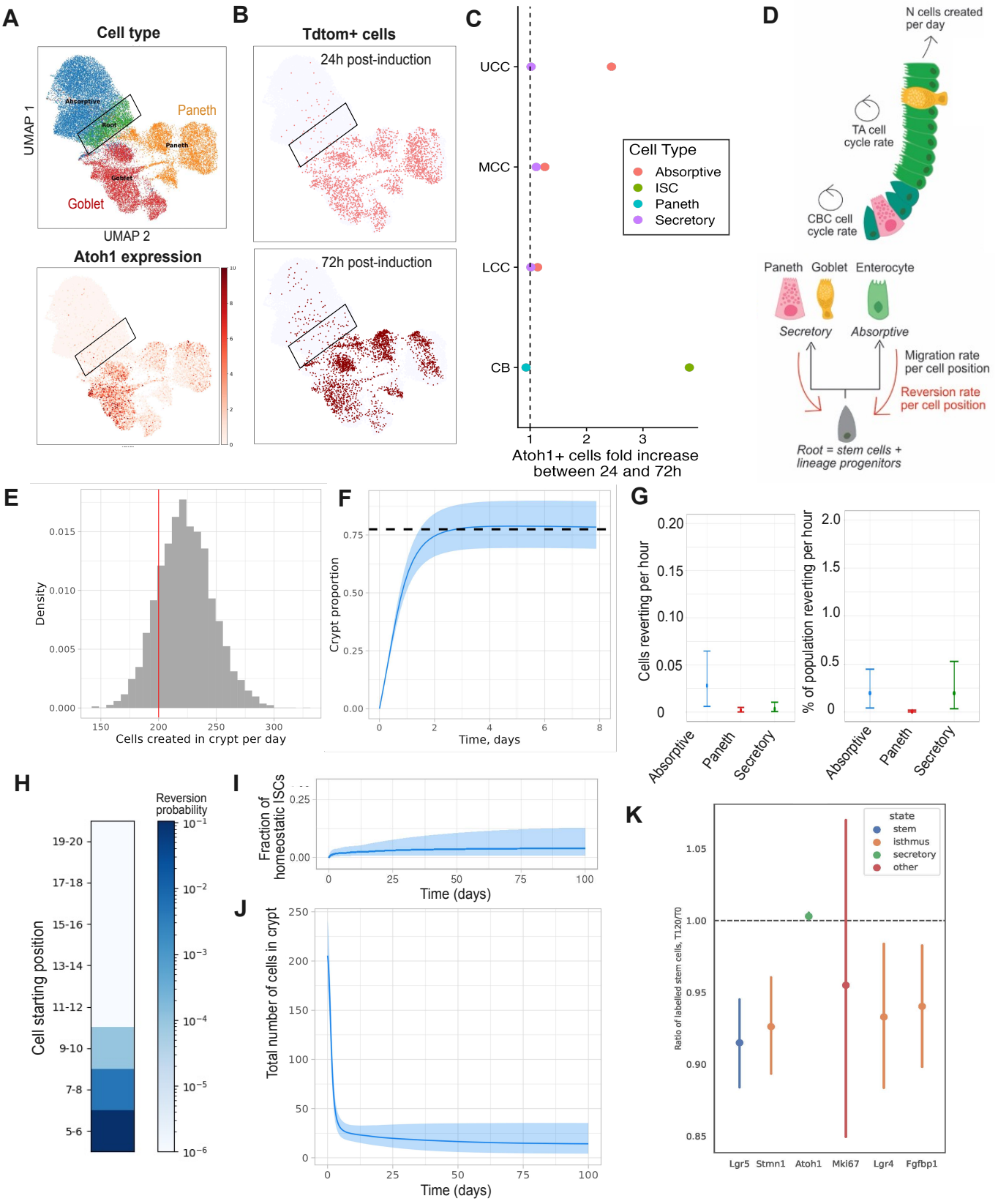


Figure 2. Homeostatic population dynamics derived from single cell data. A) UMAPs showing cell types and representative expression of Atoh1 in wild-type samples. B) UMAPs showing the location of tdTom+ cells 24h (Top)and 72h (top) after the start of tamoxifen induction. Box identifies the root (ISC) population shown in A. C) Plot showing the fold increase in proportion of Atoh1+ cells within each cell type and at each crypt region from 24h to 72h. D) Schematic representing the integration of modalities derived with either CRISP (single-cell data) or pulse chase with CldU and EdU into a population dynamics model that predicts rates of cell migration and reversion. E) Population dynamics model estimate of the number of cells created per crypt per day, the red line indicates the expected number. F) In silico lineage tracing plot of the time taken for all crypt cells (excepting Paneth cells) to acquire label from ISCs due to cell turnover. G) Dot plots showing the number of cells reverting per day for eachlineage in homeostasis (left) and the same values normalised by cell population size (right). H) Heatmap showing the probability of cells at different crypt positions reverting in the crypt base. I) Plot showing estimated ISC restoration by homeostatic dynamics after in silico ablation of stem cells. The blue line and shaded regions indicate number of ISCs generated with confidence intervals (excluding Paneth cells). J) Plot showing estimated number of cells in crypts predicted by homeostatic dynamics, after in silico ablation of ISCs. K) Ratio of labelled stem cells at T=120h post labelling compared to inital labelling for several isthmus genes, the ISC marker Lgr5, secretory marker Atoh1 and cell cycle gene Mik67.

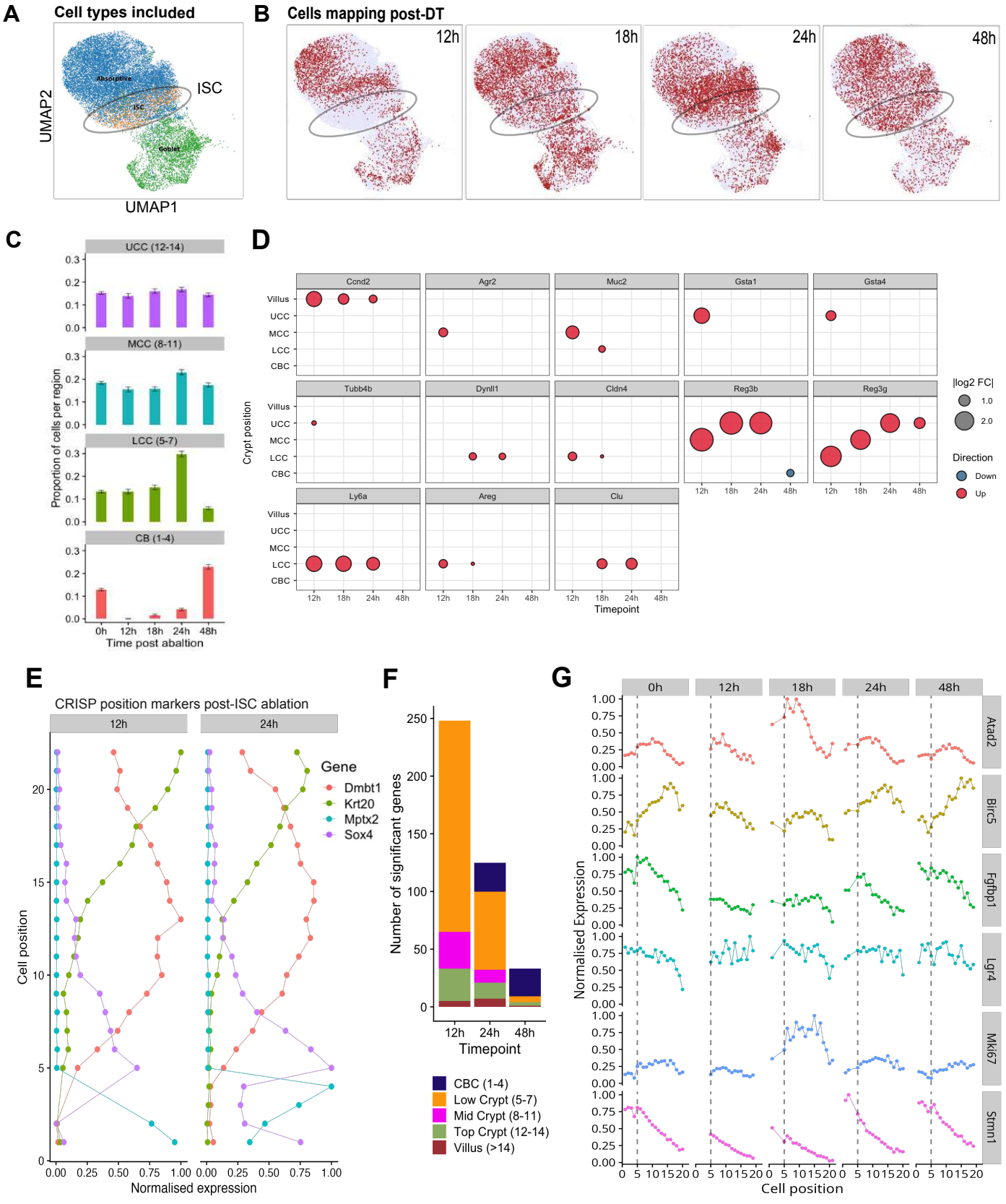


Figure 3: Single-cell spatial and proliferative mapping of regeneration after ISC ablation. A) UMAP showing cell types included in the ablation analysis. B) UMAPs mapping of the cells present at different timepoints post-ablation. C) Bar plot showing the proportion of cells per crypt region at different timepoints post-ablation. D) Bubble plot showing the maximum absolute log2 fold change (bubble size) of selected genes across crypt positions and timepoints post-ISC ablation. Only statistically significant changes are shown (p value < 0.05). E) Average expression of spatially restricted genes assigned by CRISP at 12h and 24h post-ablation, showing similar patterns compared to homeostasis. F) Barplot showing the proportion of differentially expressed genes at different timepoints post-ISC ablation and different crypt regions. G) Plot showing normalised gene expression depending on cell positions at different timepoints post-ISC ablation for the isthmus genes *Atad2, Birc5, Fgfbp1, Lgr4* and *Stmn1* as well as the proliferation marker *Mki67*. for *Ly6a* showing elevated expression in cells directly above the crypt base at 12h after ISC ablation.

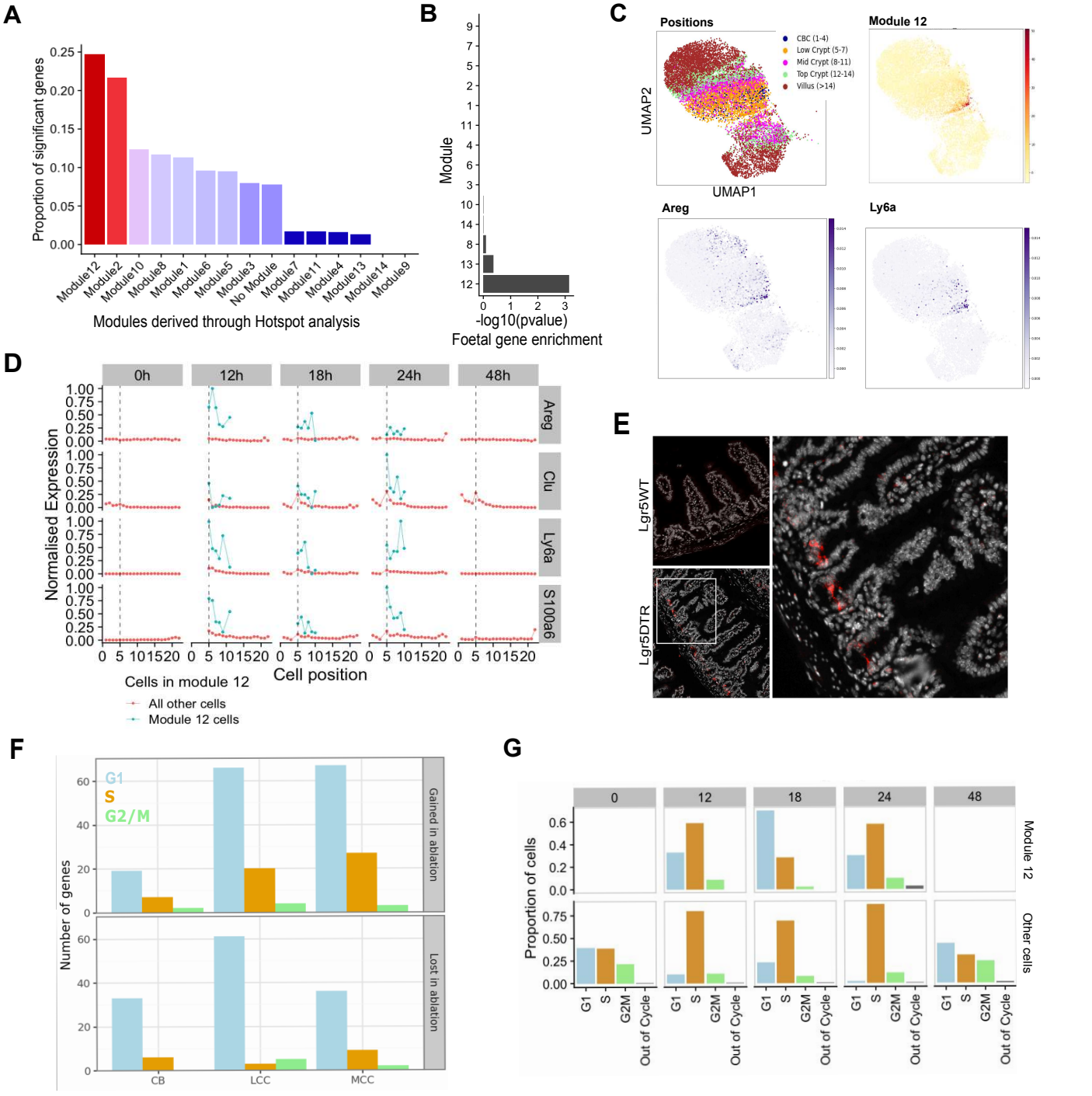


Figure 4: Identification of responsive sub-population by gene co-expression. A) Bar plot showing the proportion of genes significantly differentially expressed at 12h after ablation in each module detected with Hotspot. B) Bar plot showing foetal gene enrichment across co-expression modules, ordered by significance. Bar height represents −log10(p-value) from a one-sided Fisher's exact test (foetal dataset from Mustata et al., 2013). C) UMAP projection showing enrichment for Hotspot module 12, as well as expression of *Ly6a* and *Areg* at 12h post-ablation. D) Dot plot showing the normalised expression of foetal genes at different cell positions and timepoints post-ISC ablation in module 12 cells (blue) compared to all other cells (red). E) RNAscope RNAscope for *Ly6a* showing elevated expression in cells directly above the crypt base at 12h after ISC ablation. F) Number of genes showing differential expression during the cell cycle that are uniquely cell cycle regulated in the crypt population shown. G) Bar plot showing proportions of Module 12 and LCCs (excluding Module 12 cells) in different stages of the cell cycle.

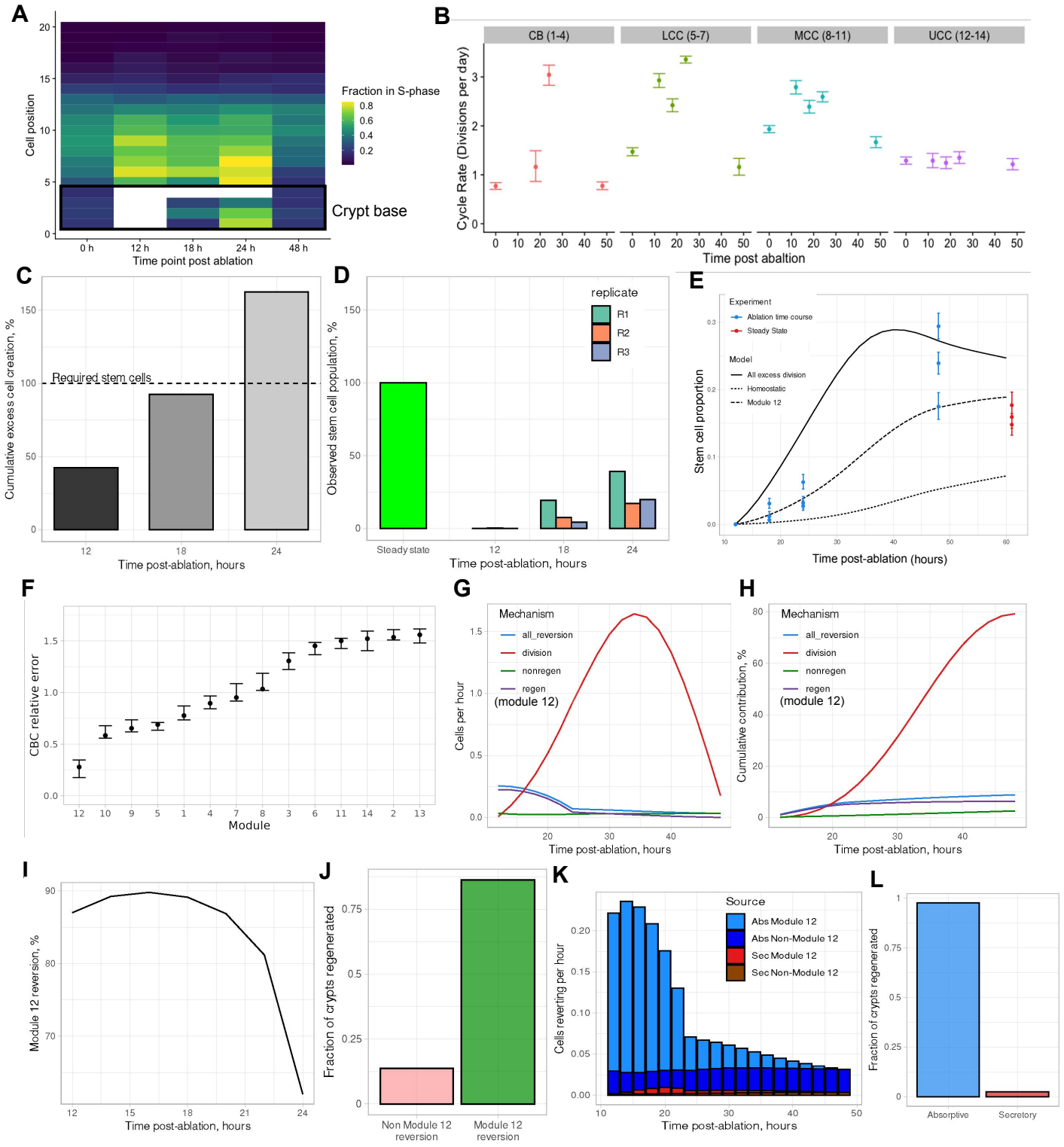


Figure 5: Population dynamics model pinpoints the source of ISC regeneration post-ablation. A) Heatmap showing the fraction of cells in S-phase at each cell position at different timepoints post-ISC ablation. B) Plot showing cycle rate for cells in different crypt regions at different timepoints post-ISC ablation. C) Bar plot showing the cumulative excess cell creation above that of the crypt base at steady state (green bar) at different timepoints post-ISC ablation (grey dotted line highlights the size of the ISC compartment at steady state). D) Frequency of ISCs observed following ablation. E) Line plot showing model prediction for the restoration of the ISCs population with time assuming it is mediated by: All excess divisions within the LCC population (solid line); the homeostatic rate of reversion rescaled by the excess cell divisions within the LCCs; Module 12 cells. Recorded stem cell proportions are shown with dots and error bars (red=steady state, blue=timepoints post-ISC ablation). F) Relative error in the population dynamics model fit when using each of the Hotspot modules as the regenerative population driving ISC restoration. G) Line plots showing model predictions for the rate of ISC restorations when originating from: Total reversion from the LCC population (blueline); Module 12 cell reversion; (purple line); homeostatic reversion rescaled by excess divisions in LCCs (green line); division of ISCs (red line). H) Line plots showing the cumulative proportion of cells reverting to the crypt base with modelling conditions as in G. I) Line plot showing model prediction for contribution of module 12 cells reversion to total restored ISCs at different timepoints post ablation. J) Stacked bar plot showing the number of cells from different branches reverting per hour at different timepoints post-ISC ablation. K,L) Bar plots showing number of crypts regenerated per 100,000 by: absorptive or secretory cells (K); by homeostatic reversion or the Module 12 cells (L).

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Supptables.xlsx
- Supp1.pdf
- Supp2.pdf
- Supp3.pdf
- Supp4.pdf
- Supp5.pdf