Charting the Single-Cell Landscape of Colorectal Cancer Stem Cell Polarisation

Ferran Cardoso Rodriguez

A dissertation submitted in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

of

University College London.

Department of Oncology

University College London

September 26, 2023

I, Ferran Cardoso Rodriguez, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the work.

Abstract

Colonic epithelia is regulated by cell-intrinsic and cell-extrinsic cues, both in homeostatic tissues and colorectal cancer (CRC), where the tumour microenvironment closely interacts with mutated epithelia. Our understanding on how these cues polarise colonic stem cell (CSC) states remains incomplete. Indeed, charting the interaction between intrinsic and stromal cues requires a systematic study yet to be found in the literature.

In this work I present my efforts towards computationally studying colonic stem cell polarisation at single-cell resolution. Leveraging the scalability of organoid models, my colleagues and I dissected the heterocellular CRC organoid system presented in Qin & Cardoso Rodriguez *et al.* [1] using single-cell *omic* analyses, resolving complex interaction and polarisation processes.

First, I identified bottlenecks in common mass cytometry (MC) analysis workflows benefiting from either increased accessibility or automation; designing the CyGNAL pipeline and developing a cell-state classifier to tackle these points respectively. I then used single-cell RNA sequencing (scRNA-seq) data to reveal a shared landscape of CSC polarisation; wherein stromal cues polarise the epithelia towards slow-cycling revival CSC (revCSC) and oncogenic mutations trap cells in a hyper-proliferative CSC (proCSC) state. I then developed a method to visualise single-cell differentation using a novel valley-ridge (VR) score, which can generate data-driven Waddington-like landscapes that recapitulate differentiation dynamics of the colonic epithelia. Finally, I explored an approach for holistic inter- and intracellular communication analysis by incorporating literature information as a directed knowledge graph (KG), showing that low-dimensional representations of the graph retain biological information and that projected cellular profiles recapitulate their transcriptomes.

These results reveal a polarisation landscape where CRC epithelia is trapped in a proCSC state refractory to stromal cues, and broadly show the importance of joint collaborative wet- and dry-lab work; central towards targeting gaps in the method space and generating a comprehensive analysis of heterocellular signalling in cancer.

Impact Statement

By investigating the intricate interplay between intrinsic and extrinsic cues regulating CSC fates, the research and work presented in this thesis sheds new light on the landscape of colonic epithelia polarisation and offers insights into potential therapeutic strategies.

Furthering the spirit of shared scientific knowledge and collaborative research, data and code used to generate the analyses in Qin & Cardoso Rodriguez *et al.* have been made public in various repositories. Furthermore, tools and outputs developed during my project and presented in this thesis have also been made publicly available; either as part of peer-reviewed publications such as CyGNAL in Sufi & Qin *et al.* [2], as software packages like pyKrack (Appendix A, `ferranc96.github.io/pyKrack`), or in the form of publicly accessible GitHub repositories.

Tools like CyGNAL and the VR landscapes have already impacted research in my lab, facilitating routine MC analyses and empowering Ramos Zapatero & Tong *et al.* [3] during the ongoing revision process. Furthermore, general knowledge acquired before and during my PhD has been shared with colleagues; either in the form of scientific discussions, empowering others to further their own technical skills, or as natural peer-peer diffusion of soft skills and life experiences.

Finally, the scientific findings shown here and in Qin & Cardoso Rodriguez *et al.* [1] will inspire and empower others in their work.

Acknowledgements

I must thank Chris for his exceptional guidance and support. A PhD is a journey of learning, and he has been the mentor whose insightful observations have been instrumental in shaping the direction of this work. I am immensely grateful to Xiao for her contributions to my PhD; from tiny details like sharing the template for my first poster, to undertaking all experimental work our analyses are based on. Her dedication and talent prove she is the best colleague one could ever work with and she is not alone. The Tape lab is a nurturing and inspiring environment full of amazing individuals; I will always remember the 5k-f^2 project and I truly hope one day our work makes a lasting impact on patients' lives.

I am also grateful to the UCL-Yale exchange programme, for without their bursary I would not have been able to collaborate with Smita and her team of exceptional individuals. Thank you Aarthi for your unwavering support despite the setbacks faced, and thank you too Jay for your help during our discussions of KGs and hikes around Connecticut. I would also like to thank Jasmin, Nicky and Javier for their guidance as my thesis committee panel; thank you Javier and other BLIC members for your support and insightful discussions too. I must thank CRUK and similar organisation supporting scientific research; without your help none of this would be possible.

Thank you Mum and Dad, and you too Sara; I would not be where I am nor who I am without you. *Moltes gràcies*, this thesis goes to you. Last but not least, I must thank you too Ana; my soon-to-be wife and anchor throughout this journey. Thank you for sharing of the highs and lows of academia, and for sharing a life with me. Also, I am sure this will not be the last thesis acknowledging you. *T'estimo*.

UCL Research Paper Declaration Form: Chapter 3

1. **1. For a research manuscript that has already been published** (if not yet published, please skip to section 2)**:**

   (a) **What is the title of the manuscript?** Multiplexed Single-cell Analysis of Organoid Signaling Networks.

   (b) **Please include a link to or doi for the work:** `https://doi.org/10.1038/s41596-021-00603-4`

   (c) **Where was the work published?** Nature Protocols.

   (d) **Who published the work?** Springer Nature.

   (e) **When was the work published?** 08 September 2021.

   (f) **List the manuscript's authors in the order they appear on the publication:** Jahangir Sufi, Xiao Qin, Ferran Cardoso Rodriguez, Yong Jia Bu, Petra Vlckova, María Ramos Zapatero, Mark Nitz, and Christopher J. Tape.

   (g) **Was the work peer reviewed?** Yes.

   (h) **Have you retained the copyright?** Yes.

   (i) **Was an earlier form of the manuscript uploaded to a preprint server (e.g. medRxiv)? If 'Yes', please give a link or doi** No.

2. **For a research manuscript prepared for publication but that has not yet been published** (if already published, please skip to section 3)**:**

3. **For multi-authored work, please give a statement of contribution covering all authors** (if single-author, please skip to section 4)**:** J.S. developed TOB*is*, designed rare earth metal-conjugated antibody panels and performed MC analysis. X.Q. designed and performed organoid and MC experiments, analysed the data and wrote the manuscript. F.C.R. developed CyGNAL and wrote the manuscript. P.V. and M.R.Z. performed organoid and MC experiments. Y.J.B. and M.N. developed TeMal reagents. C.J.T. designed the study, analysed the data and wrote the manuscript.

4. **In which chapter(s) of your thesis can this material be found?** Chapter 2 (Methods) and Chapter 3.

Signatures confirming that the information above is accurate

Candidate: Ferran Cardoso Rodriguez

Date: 26 July 2023

Supervisor/Senior Author signature (where appropriate)**:** Christopher J. Tape

Date: 26 July 2023

UCL Research Paper Declaration Form: Chapters 4-5

1. **1. For a research manuscript that has already been published** (if not yet published, please skip to section 2)**:**
2. **For a research manuscript prepared for publication but that has not yet been published** (if already published, please skip to section 3)**:**
   (a) **What is the current title of the manuscript?** A Single-cell Perturbation Landscape of Colonic Stem Cell Polarisation.
   (b) **Has the manuscript been uploaded to a preprint server 'e.g. medRxiv'?** Yes.
   If 'Yes', please please give a link or doi: `https://doi.org/10.1101/2023.02.15.528008`
   (c) **Where is the work intended to be published?** Cell
   (d) **List the manuscript's authors in the intended authorship order:** Xiao Qin*, Ferran Cardoso Rodriguez*, Jahangir Sufi, Petra Vlckova, Jeroen Claus, and Christopher J. Tape. *: These authors contributed equally to this work.
   (e) **Stage of publication:** In revision, resubmitted.
3. **For multi-authored work, please give a statement of contribution covering all authors** (if single-author, please skip to section 4)**:** X.Q. designed the study, performed organoid experiments, generated scRNA-seq and TOB*is* mass cytometry data, analysed mass cytometry data, and wrote the paper. F.C.R. analysed scRNA-seq data, developed the VR score, and wrote the paper. J.S. developed TOB*is* barcodes and conjugated rare-earth metal antibodies. P.V. provided organoid culture support. J.C rendered Waddington-like landscapes. C.J.T. designed the study, analysed the data, and wrote the paper.
4. **In which chapter(s) of your thesis can this material be found?** Chapter 2 (Methods) and Chapters 4 and 5.

Signatures confirming that the information above is accurate

Candidate: Ferran Cardoso Rodriguez

Date: 26 July 2023

Supervisor/Senior Author signature (where appropriate)**:** Christopher J. Tape

Date: 26 July 2023

Abbreviations

k-NN k-Nearest Neighbours. 18, 42–46, 59, 62, 63

A *shApc*. 19, 79, 92, 96, 97, 100

AK *shApc* and $Kras^{G12D/+}$. 19, 21, 22, 79, 92, 96–98, 100, 103, 104

AKP *shApc*, $Kras^{G12D/+}$ and $Trp53^{R172H/-}$. 19, 21, 79, 92, 96, 97, 100

CRC colorectal cancer. 3, 17, 22, 28, 30, 31, 47, 75, 104, 127, 130

CSC colonic stem cell. 3, 4, 20, 21, 23, 29–31, 94, 96, 112, 127, 128

DA differential abundance / differentially abundant. 21, 62, 96, 97

DCS deep crypt secretory. 17, 30

DE differential expression / differentially expressed. 21, 44, 61, 62, 97

DR dimensionality reduction. 41, 126

DREMI Density Resampled Estimate of Mutual Information. 18, 76, 77, 80, 88, 125

EEC enteroendocrine. 17, 29, 30

EMD Earth Mover's Distance. 18, 76, 77, 80, 88, 125

ER endoplasmic reticulum. 17, 29, 30

GEx gene expression. 21, 23, 24, 26, 97, 121–124, 131, 151

KG knowledge graph. 3, 5, 18, 23, 25, 46, 47, 69–72, 115–120, 123, 131, 132

LRT-KG ligand-receptor-target KG. 23, 24, 26, 117, 118, 121, 123, 131, 152

MC mass cytometry. 3, 4, 18, 19, 34, 35, 47, 75–77, 79, 80, 83, 88, 105, 121, 125–128

NGS next-generation sequencing. 35, 37

PCA Principal Component Analysis. 18, 41, 77, 88

PDO patient-derived organoid. 20, 32, 86

proCSC hyper-proliferative CSC. 3, 20, 30, 31, 91, 94, 96, 105, 106, 111, 113, 127–130

PTM post-translational modification. 23, 34, 41, 75–77, 80, 83, 86, 115, 118, 125, 126, 131, 132

QC quality control. 40

revCSC revival CSC. 3, 20, 30, 31, 91, 94, 105, 106, 111, 113, 121, 122, 127–131

RF Random Forest. 19, 20, 75, 84–86, 88, 89, 126

scRNA-seq single-cell RNA sequencing. 3, 20, 23, 34, 35, 80, 92, 105, 121, 127, 131

TA transit amplifying. 17, 20, 29, 30, 94, 98, 122

TF transcription-factor. 23, 69, 71, 115, 117, 118

TME tumour microenvironment. 28, 31, 47, 75

TOB*is* Thiol Organoid Barcoding *in situ*. 6, 8, 75, 80, 83

UMAP Uniform Manifold Approximation and Projection. 18, 77–79

UMI unique molecular identifier. 36

VR valley-ridge. 3, 4, 18, 22, 23, 68, 69, 109, 110, 112, 127, 129, 130, 136

WT wild-type. 19, 21–24, 26, 79, 84, 92, 96–98, 100, 103, 104, 121, 128, 131, 138, 151, 152

Contents

List of Figures

List of Tables

Chapter 1

Introduction and Background

1.1 Significance and Characteristics of Colorectal Cancer

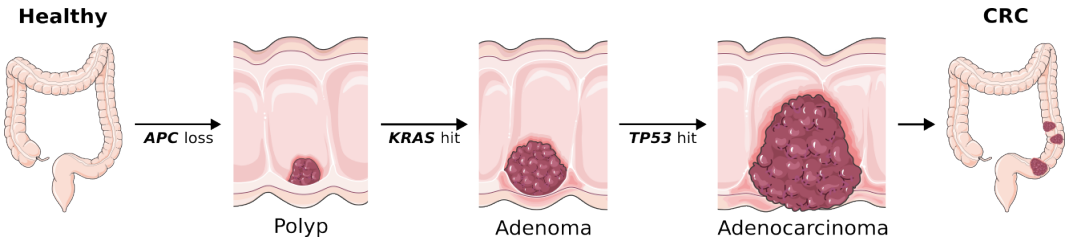


Figure 1.1: Canonical Progression Model of CRC. Schematic diagram depicting the transition from healthy colon, to adenoma, adenocarcinoma, and CRC via the accumulation of cell-intrinsic somatic oncogenic mutations.

Colorectal cancer (CRC) is generally defined as an adenocarcinoma originating from the epithelial lining of the colon or rectum. Despite lowered incidence and mortality rates in recent years [10], CRC is the third most common malignancy worldwide, claiming over 900,000 lives every year [11].

The canonical model of CRC pathogenesis is the polyp to adenocarcinoma progression. Originally described at the end of the 20th century by *Fearon and Vogelstein* [12] it is understood to present an initial phase where benign hyperproliferative polyps, often harbouring mutations in the *Wnt* signalling pathway (most commonly in the *APC* gene [13]), eventually acquire additional oncogenic mutations that result in malignant CRC (Figure 1.1). Some of the most common oncogenic mutations target *KRAS*, an oncogene that regulates epithelial proliferation, and *TP53*, a tumour suppressor that normally acts as a gatekeeper of the hyperproliferative polyps [12, 14].

Furthermore, the development of CRC also involves the local tumour microenvironment (TME), whereby the mutated epithelial cells orchestrate changes in the local inflammatory and stromal niches [15].

1.1.1 The Colonic Epithelium and its Stem Cells

The intestinal epithelium comprises an epithelial mono-layer lining the lower gastrointestinal tract that controls nutrient uptake, coordinates metabolism, and shields against pathogens. In a homeostatic setting, intestinal epithelia has an extremely

high turnover rate and is organised as distinct cell populations with absorptive or secretory functions, supported by continuously proliferating crypts [16]. The colon and rectum form the distal end of the gastrointestinal tract and, unlike the longer small intestinal compartment, experience a higher microbial load, lack villi, and specialise in liquid uptake [17].

At the base of the colonic crypts reside $LGR5^+$ and $OLFM4^+$ colonic stem cells (CSCs) that give rise to rapidly proliferating transit amplifying (TA) cells (Figure 1.2A). While the specific differentiation trajectories are not yet fully understood, it seems that an endoplasmic reticulum (ER) stress response marks the shift from a basal proliferation state into differentiated epithelial states (Figure 1.2A) [18, 19]. Of those differentiated states the most common ones are the enterocytes with an absorptive (also called colonocytes in the colon), and secretory cells such as; mucus-secreting goblet cells, hormone-producing enteroendocrine (EEC) cells, and immuno-modulatory tuft cells.

The delicate balance of spatial and temporal control of cell fate is achieved by two opposing gradients between the basal and apical folds of the epithelium, with WNT and NOTCH signalling higher around the CSC-harbouring crypts, and BMP signalling higher towards the apical areas where absorptive cells are (Figure 1.2A) [16, 20]. Continued epithelial renewal is sustained by the CSC population. Characterised by their expression of the LGR5 R-spondin receptor, CSCs are primed to receive converging signalling cues from stromal and intrinsic signals that delineate areas of cell differentiation and proliferation.

Although this arrangement is kept relatively consistent throughout the lower gastrointestinal tract, organoid models suggest that the architecture of the homeostatic crypt in the colon appears to be, unlike that of the small intestine, more dependent on exogenous stroma-derived WNT ligands and BMP antagonists [21, 22]. This difference is thought to be driven by secretory cells known as Paneth cells, which reside at the bottom of the crypts in the small intestine but are absent in the colon. Paneth cells support nearby stem cells through the secretion of antimicrobial peptides, WNT and EGF ligands, and juxtacrine NOTCH signalling. In the colon the presence

of secretory cells in deeper areas of the crypts has been described [23], but it is believed that the niche supporting the stem compartment is mostly orchestrated by the stroma rather than by these Paneth-like deep crypt secretory (DCS) cells.

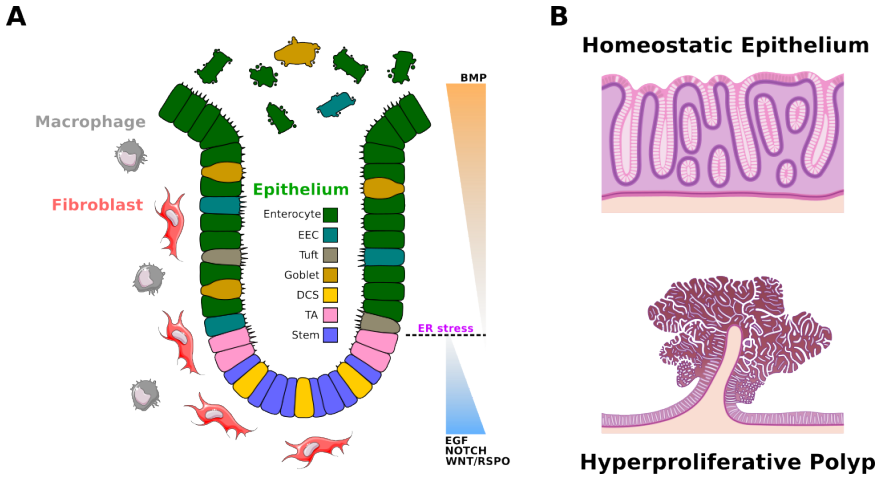


Figure 1.2: Architecture of Colonic Epithelium. **A)** Colonic cell types and signalling gradients regulating stem and differentiated epithelial niches. **B)** Tissue morphology schematic of homeostatic colonic epithelium and hyperproliferative colonic polyps. DCS, deep crypt secretory. EEC, enteroendocrine. ER, endoplasmic reticulum. TA, transit amplifying.

1.1.2 Colorectal Cancer as a Heterocellular Disease

Genetic alterations in epithelial cells commonly target niche factor signalling hubs that regulate proliferation and differentiation, enabling the CSC compartment to decouple from both pro-survival proliferative signals and growth-inhibitory cues [24]. This results in an emancipated and highly proliferative stem-like state (proCSC) that expands beyond the bases of the crypts and dominates the colon epithelium, thus accompanied by a general de-differentiation of the tissue [25] (Figure 1.2B).

Although it is tempting to think that the expansion of the CSC compartment in CRC is driven by this highly proliferative homogeneous proCSC state, single-cell studies have revealed the presence of additional stem cell states in both homeostatic and CRC epithelium [26, 27, 28, 29]. Among them, revival CSCs (revCSC) are emerging as a target of particular interest in cancer research. A rare population in the homeostatic intestine, revCSCs are characterised by *CLU* and *ANXA1* expression and exhibit a less proliferative state that, upon tissue damage, co-opts a phenotype

reminiscent of foetal intestinal progenitors to replenish the injured epithelium [30]. In the context of CRC, revCSC have been postulated as a putative drug-resistant state that can, after chemotherapy erodes the dominant proCSC state, drive relapse in some patients [31, 32]. While the revCSC state has been associated with Hippo pathway activation via YAP signalling, their exact role in relapse and the mechanisms driving the balance between revival and proliferative CSCs remain unclear.

A priori a niche-factor independent compartment, the CRC epithelium comprised mostly of emancipated CSC and proCSC cells is still able to interact and remodel surrounding tissues. This interaction with their environment sustains the view that tumours exist not just as homogeneous clusters of malignant cells, but as a collection of malignant and non-transformed immune and stromal cells [33]. These untransformed cells constitute the tumour microenvironment (TME), a key factor in most cancers that affects prognosis [34] and therefore the subject of intense study in cancer biology and therapy development.

In their late stage, CRC tumours consist of a complex heterocellular environment in which stromal and immune compartments have been shown to drive cancer cell progression [15, 35] and response to therapies [36, 37]. Cancer associated fibroblasts in particular have been linked with carcinogenesis via secretion of growth factors like EGF, HGF, VEGF and TGF-β signalling. In addition, they have also been linked with pro-inflammatory and angiogenic roles, as well as with aiding the CRC tumour in immune evasion and invasion [38]. Within the immune compartment, tumour-associated macrophages are highly abundant, but their functional role as part of the TME is unclear. There is evidence that they both exhibit pro- and anti-tumour activity, possibly depending on their location within the adenocarcinoma and the dominance of different macrophage sub-types [39].

1.2 Organoids as *In Vitro* Models of Colorectal Cancer

The complexity of CRC can be modelled and studied *in vitro* using organoids, self-organising 3D cellular structures comprising stem and differentiated cells that mimic elements of *in vivo* tissue [40, 41, 42]. Mimicking the biology of the *in vivo* setting, gut organoids have a basal stem niche from which differentiated states (with absorptive or secretory functions) derive from; often with an apical lumen within the organoid that accumulates dead cells [43].

Furthermore, heterotypic settings can be designed wherein colon epithelia organoids are co-cultured with other cell types to model stromal and immune cell-cell interactions [4]. Such settings increase the complexity of organoid systems, allowing for more accurate modelling of *in vivo* tissue architecture and heterotypic interactions *in vitro*.

In the context of CRC, organoids can be used to characterise both the heterogeneity of the altered colonic epithelium and its interaction with cells of the TME. Furthermore, patient-derived organoid (PDO) models are gaining traction as personalised avatars of human tumours [44, 3].

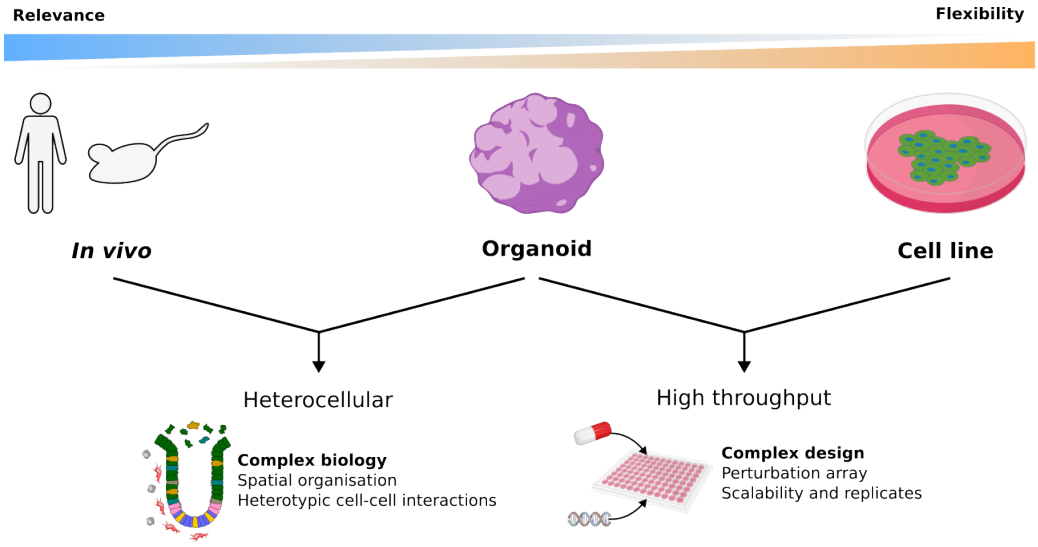


Figure 1.3: Organoids as CRC Models. Organoids balance the physiological relevance of *in vivo* settings with the flexibility of cell lines, allowing for high-throughput study of complex heterocellular systems such as CRC.

Organoids provide a balance between experimental flexibility and physiological relevance. They are complex enough to mimic the heterogeneity of *in vivo* tissue while still being amenable to high-throughput applications [45]. This facilitates high throughput experimentation by allowing for the multiplexing of high numbers of experimental conditions, with for example our custom mass cytometry platform in Sufi & Qin *et al.* reaching up to 126-plex per run [2].

Recent work by our lab [4] has shown how both CRC genetic perturbations (*shApc*, $Kras^{G12D/+}$ and $Trp53^{R172H/-}$) and TME complexity (heterotypic epithelial organoid cultures with fibroblasts and/or macrophages) effect the biology of colonic organoids. Using a custom multivariate mass cytometry platform to analyse post-translational modification (PTM) signalling networks, Qin *et al.* [4] found that the distribution of both cellular subtypes and states within the epithelial population changed in a similar and synergic way. They found that both oncogenic and stromal cues resulted in an enrichment of the crypt and stem niches and a reduction of cells in G0 and apoptotic states. Furthermore, their results suggest that the effects of the TME on intracellular epithelial signalling pathways might mechanistically differ from those driven by CRC mutations in the epithelial cells, even if they both share downstream signalling profiles.

This work, featuring multiple axes of variation and replicates within a single experiment, highlights the systematic scalability of organoid models. Mature bulk technologies are not poised to leverage heterogeneous 3D organoids, hence the rapid emergence of single-cell resolution studies in recent years. Single-cell *omic* approaches can deconvolute the different cell types within a heterotypic organoid system, as well as resolve particular cell states within each type and even capture cellular interactions within the different compartments [36].

1.3 Single-Cell *Omic* Technologies

During this work I leveraged two distinct single-cell technologies to characterise heterocellular organoid models of CRC; mass cytometry (MC) and single-cell RNA sequencing (scRNA-seq). They are both part of the broader family of single-cell *omics* analyses, which have gained traction in characterising cellular heterogeneity at both genotypic and phenotypic levels.

The concept of *"omics"* is not well defined, but it is commonly understood to describe analyses pertaining to the study of large-scale biological datasets characterising sets of biological molecules from living entities. Some of the most common *omic* studies are the fields of genomics, epigenomics, transcriptomics, and proteomics. *Omic* information can thus be used to infer cross-*omic* regulatory relationships and decipher causal relations between genotype and phenotype with the right experimental settings.

1.3.1 Mass Cytometry (MC)

MC, also known as Cytometry by Time-Of-Flight (CyTOF), is a technology that merges principles of mass spectometry and flow cytometry to enable single-cell analysis of protein expression. Like flow cytometry, MC is based on tagged antibodies that bind to specific epitopes in cells, but it is able to overcome the issue of florescent spectral overlap by using monoisotopic rare-earth metals instead of fluorophores. The discrete nature of the monoisotopes compared to the broad emission spectra of fluorophores allows for the design of antibody panels that can capture up to $1 \cdot 10^2$ features per cell [46].

Resolving total protein level information in single-cells is in itself incredibly useful, but MC also excels at resolving post-translational modifications (PTMs) [47]. PTM information often determines a cell's state in relation to the cell cycle, as this process is not really regulated at the gene level but rather by a tight control of different PTM-driven checkpoints [48]. This capability also allows for in-depth study of intracellular signalling networks, DNA-damage responses, and apoptosis; having already been used to characterise both cell-state and oncogene- and stroma-driven signalling changes in murine CRC organoid models [4].

Coupled with a custom multiplexing platform [2] MC technology can analyse extremely wide experimental systems covering a large number of conditions and replicates, which proves especially useful for drug screening applications [3].

However, while powerful in the study of intracellular signalling, mass cytometry struggles to resolve intercellular communication through the complex extracellular interactome of ligands and receptors. In contrast, single-cell RNA sequencing technologies can prove extremely useful for this purpose, especially when combined with intercellular cell communication databases such as CellChat [49] and CellPhoneDB [50].

1.3.2 Single-Cell RNA Sequencing (scRNA-seq)

With the advent of next-generation sequencing (NGS) technologies, bulk-based RNA sequencing approaches were devised that could capture genome-wide transcriptomic information from a whole sample. This mature technology enabled key discoveries across a variety of fields, including tissue development and cancer biology, but its inability to resolve individual cells and their states is a key limitation in systems with complex transcriptional dynamics and multiple cell types [51]. scRNA-seq overcame this issue by capturing transcriptomic information at the level of individual cells. Now, a collection of discrete transcriptomic profiles can be pieced together to recapitulate continuous differentiation trajectories, or complex heterocellular systems could now be resolved into their individual cell types [52]. However, while powerful, scRNA-seq comes with significant technical challenges and costs.

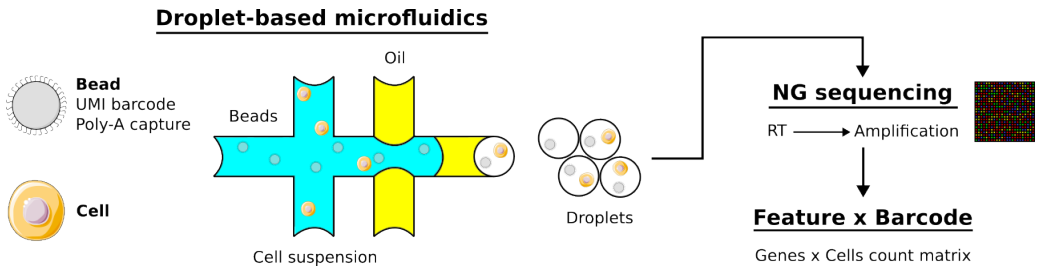


Figure 1.4: Droplet-Based Single-Cell RNA Sequencing. Microfluidics platforms barcode single-cells via cell-bead droplet encapsulation. Barcoded cells are sequenced to generate count matrices. NG, next-generation (sequencing). RT, reverse transcription. UMI, unique molecular identifier.

Mature and highly optimised microfluidic droplet-based approaches tend to dominate the commercial market, with 10X Genomics offering commercial products [53] that perform the best in terms of UMI and gene / cell detection whilst being a high-throughput application [54].

Droplet-based scRNA-seq methods work by encapsulating individual cells and uniquely tagged beads into water-in-oil droplets, where the cells and beads constitute the dispersed phase and the oil forms the continuous phase encapsulating the droplets [9] (Figure 1.4). During amplification using the poly-A tail capture primers (Figure 1.4), a unique cellular barcode is added and shared across all products from a single droplet, and a unique molecular identifier (UMI) is also added as a transcript-specific tag before amplification. Resolving the single-cell level data then relies on only one cell being present in each droplet, so to avoid duplicates a significant percentage of droplets are left empty [55]. scRNA-seq methods are also characterised by dropout effects, as they capture genes with relatively low yields, resulting in sparse and noisy datasets [56].

Despite their good performance and field dominance, high throughput droplet-based microfluidic scRNA-seq approaches still represent a significant monetary burden due to library preparation, which negatively affects scalability and might even, in extreme cases, jeopardise scientific validity by potentially constraining the presence or number of replicates [57].

To overcome this burden, there has been an emergence of microfluidic-free approaches in recent times. Clark *et al.* recently developed PIP-seq [58], a droplet-based approach based on vortexer emulsification that aims to reduce costs and protocol complexity. By contrast, split-pool barcoding approaches do require a considerable amount of liquid handling steps but promise incredible scalability by using combinatorial split and pooling steps to uniquely barcode at once all cells within a sample [59].

In the context of CRC, scRNA-seq has been widely used to describe intestinal epithelia *in situ* [52] and even in organoid models, but to date no systematic analysis of colon epithelia across multiple perturbation axes capturing both CRC oncogenic

status and changes in the TME has been performed.

Also known as massively parallel methods, NGS transcriptomics requires the isolation and lysis of cells, reverse transcription of their RNA into cDNA, and then amplification to generate sequencing libraries (Figure 1.4). Despite being relatively mature technologies, it is still an advancing field, with costs reduction following Moore's Law during the last decade [60]. Emerging third-generation sequencing technologies [61] are capable of sequencing at the single-molecule level and generally produce reads that are longer than those of NGS approaches [62, 63]. Able to also measure multiple *omic* layers [64], they are poised to challenge the more common NGS technologies in the future.

1.4 Single-Cell *Omic* Data Analysis

Single-cell technologies generate *omic* scale profiles at the resolution of individual cells, so that complex heterocellular systems like organoids or *in vivo* tissues can be profiled. However, these approaches produce extremely high dimensional datasets due to the large-scale nature of *omic* data and the single-cell resolution of the technology. Although the large amount of data generated certainly does present a technical challenge, it also allows for a myriad of complex analytical approaches that leverage its complexity and depth to the fullest extent [65, 45].

1.4.1 The Three Axes of Dimensionality

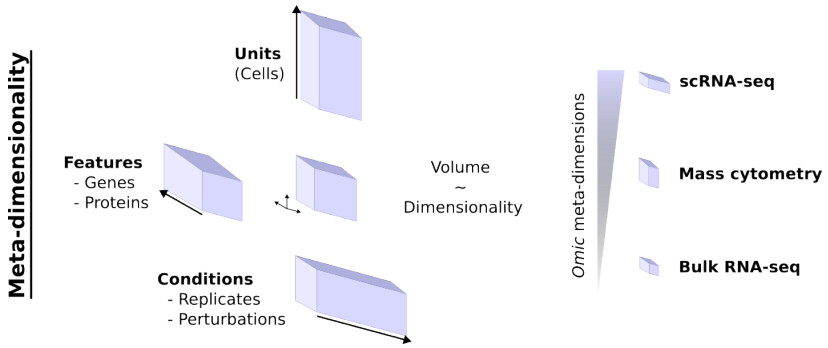


Figure 1.5: The Three Axes of Dimensionality. *Omic* data presents with unit, feature, and condition spaces that inform the concept of data meta-dimensionality.

Within the context of *omic* approaches, data dimensionality can be thought of three distinct axes; 1) the number of features to be measured, such as genes or proteins, 2) the number of units whose features are measured, and 3) the number of conditions, groups of units representing a particular biological setting [45]. Thus, a concept of meta-dimensionality is useful to refer to all axes at once. The unit of measurement is dependent on the methodology used, with bulk methods measuring at the level of whole samples whereas single-cell approaches resolve individual cells. Some spatial *omic* methods fall somewhere in between bulk and single-cell approaches, examining specific regions of a sample containing a small number of cells [66, 67, 68].

Thus, single-cell approaches can generate extremely high-dimensional datasets due to the large-scale nature of *omic* data and the single-cell resolution of the technology. This presents new data analysis challenges that are further compounded

when applied to highly scalable models such as organoids that allow for high numbers of conditions to be measured. Machine learning approaches and dimensionality reduction techniques are thus commonly applied to extract meaningful information from high dimensional single-cell data, and, while there is still no uncontested consensus, the more common approaches will be discussed below.

1.4.2 Data Integration

The essence of data integration is the merging of multiple discrete datasets, and their applications range from batch correction to disjoint cross-modality integration, where modality refers to different sets of measures generally across different *omic* fields.

Of the different types of integration tasks, the most common is between datasets with feature overlap but different units (cells) being measured. This type of integration is needed when datasets are acquired as different events, where generating a combined feature space onto which the cells are projected is relatively straightforward (if indeed necessary at all) and the goal is to remove any technical noise while conserving the biological signal. Data integration approaches range from simple linear methods like mean-centring adjustment commonly used for batch effect correction [69], to more complex approaches such as canonical correlation analysis [70] which uses shared anchors to integrate datasets with partially overlapping features. With the later having a tendency towards over-smoothing biological signals, recent methods like STACAS [71] have been proposed to integrate samples with heterogeneous cell states that might only partially overlap.

Alternatively, sometimes it is necessary to integrate across datasets joint along the cell axis but with different feature sets. A quite common occurrence when dealing with multi-modal techniques, this task can be approached in several ways. The oldest approaches attempted to map the modalities into a shared feature space using cross-omic prior knowledge [72], but these have mostly been replaced by techniques that consider the different modalities to be representations of the same underlying manifold, thus attempting to align the two spaces with techniques such as optimal transport while also optionally incorporating prior knowledge [73, 74].

Finally, the most challenging integration tasks are those in which there is no overlap between feature or unit spaces. In these cases, integration relies on the assumption that the cells analysed belong to the same underlying manifold of cell states (i.e. they are different snapshots of the same biological process being sampled), and allows for *in silico* generation of cross-omic integrated space from multiple disjoint unimodal datasets and atlases [75, 76, 74].

1.4.3 Common Practices for Data Analysis

Analysis of single-cell omic data is a growing and mostly non-standardised field where a myriad of tools and approaches have been proposed to leverage rich and high-dimensional single-cell omic datasets. Structurally, it is commonly divided between pre-processing and downstream analyses, and while there are some general guidelines and approaches pervasive to the field [77, 78], even very established tenets like the unsupervised clustering of cells continue to be debated.

Pre-processing of the data encompasses from more upstream tasks such as sequence alignment and feature normalisation, to further downstream steps like data integration (Figure 1.6A), commonly done after a certain degree of exploration of the feature and unit spaces. In the case of scRNA-seq, the first step is to align the sequenced reads against a transcriptome of reference [79, 80]. This process enables the generation of a count matrix that represents the unit X feature space, i.e. the gene expression detected for each gene (feature) on each cell (unit).

Once the $cellXfeature$ matrix has been generated, filtering-based quality control (QC) is performed, whereby cells that do not meet thresholds set on the feature space are removed. Commonly, as part of QC protocols doublet and apoptotic or otherwise compromised cells also get removed. The filtered data is then transformed and scaled to account for factors that might obscure biological signals, such as differences in cell metrics or feature detection capabilities and sequencing depth. These normalisation steps vary according to the data being analysed, so that for mass cytometry datasets intensities are usually normalised used an inverse hyperbolic sine transform ($asinhx$) with a co-factor of 5 [81, 82]. For scRNA-seq the approaches range from simpler (and seemingly more robust) log-based transformation [83] and

depth-based normalisation [84], to more complex methods like SCTransform [85] that use Generalised Linear Methods with Pearson residuals and are able to regress out unwanted sources of variation.

Feature selection is a common pre-processing step that precedes downstream analysis. In the sequencing field, feature selection is commonly limited to selecting highly variable genes, as it is assumed that those will carry relevant biological information and will also speed up compute time by limiting the large feature space. In less feature-rich *omic* technologies, such as mass cytometry, the aim of feature selection is rather a temporary process wherein certain features are used to determine a specific metric (such as nested Boolean gating of cell-cycle associated PTMs to determine cell-state [86, 4]). Often times it is done in conjunction with the normalisation steps commonly performed upstream (Figure 1.6A).

If relevant, data integration is commonly performed after the QC and normalisation steps, most commonly with the aim of either removing batch effects between samples or to generate a shared feature space across modalities [74].

Dimensionality reduction (DR) techniques aim to reduce the complexity of the the data while still preserving as much information as possible. If we consider that individual cells belong to a manifold where local structure can be mapped to an Euclidean space our aim would be to preserve distances between cells both in this local space but also at the global level across distant points in the manifold. Principal Component Analysis (PCA) [87, 88] was defined in the pre-computational era of the early 20^{th} century and is still commonly used due to its simplicity and speed. However, PCA is only capable of capturing linear relationships, and thus is generally used as an intermediate DR approach where high dimensional data is compressed to a feature space of $1 \cdot 10^1$ to $1 \cdot 10^2$. Later DR approaches aim to capture non-linear relationships and to better reflect the underlying manifold, and include methods like Diffusion Maps [89], t-SNE [90] and UMAP [91]. While these methods are able to preserve local distances from the manifold in the embedded space, in recent years there has been a push towards consistently preserving global manifold structure too. Methods like PHATE [92] and its multi-scale derivative [93] represent some of those

efforts that have been developed specifically for the field of single-cell *omic* data.

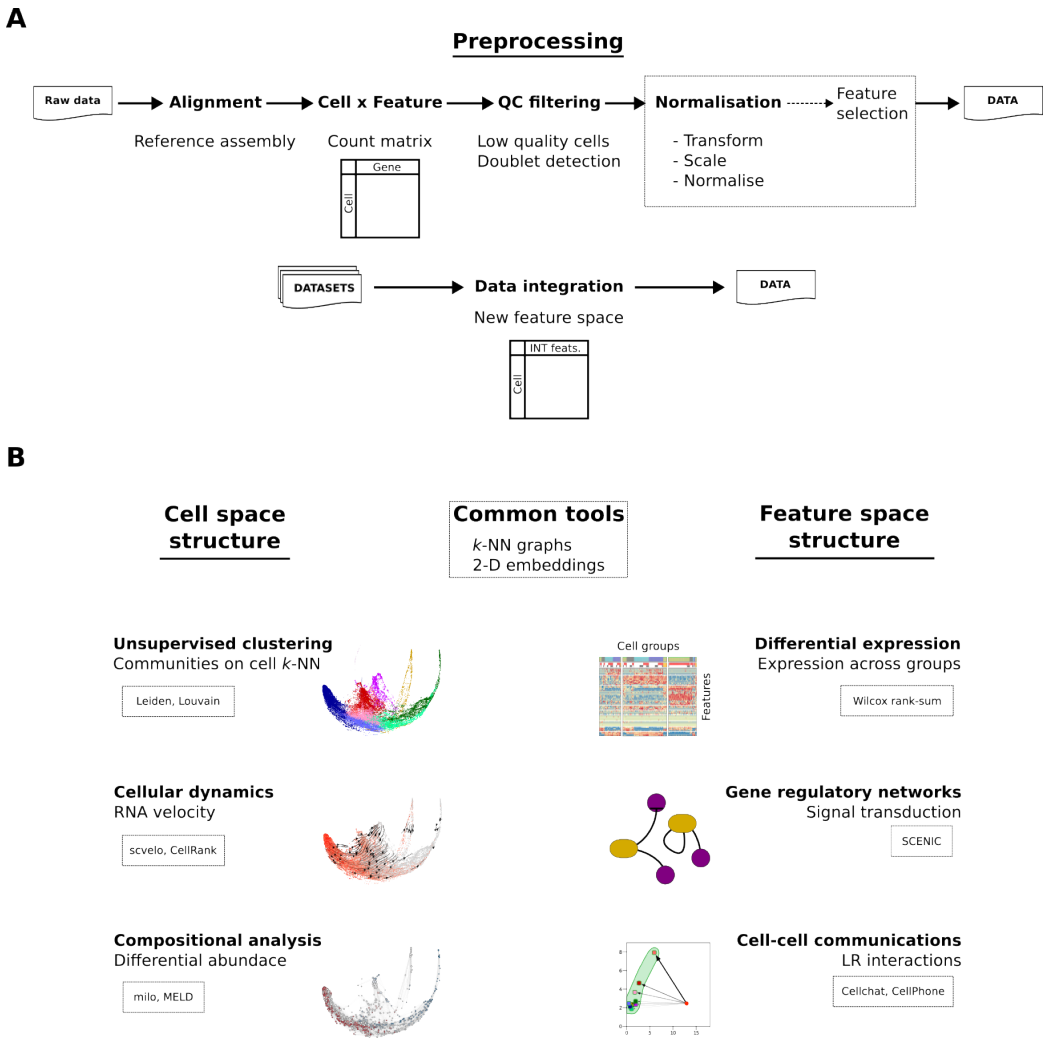


Figure 1.6: Common Practices for scRNA-Seq Data Analysis. A) Pre-processing steps encompass from alignment of sequencing data to normalisation and data integration methods. **B)** Single-cell *omic* data analysis aim to resolve structure in either the cell or feature spaces. *k*-NN, *k*-Nearest Neighbours. LR, ligand-receptor.

Downstream analyses vary in both aims and methods used, so much so that there is no uncontested gold standard data analysis workflow [94]. Despite this, all approaches tend to share a common purpose in finding structure in either the feature space (i.e. genes in scRNA-seq) or the cell space (Figure 1.6B).

Determining structure in the unit (cell) space generally translates to identifying a cell's state or type. This is commonly accomplished via unsupervised clustering methods that group cells based on their location within a *k*-Nearest Neighbours

(k-NN) graph that captures their transcriptional similarities. The output of these community detection algorithms [95, 96] is indeed unsupervised, but clusters are most commonly presented as annotated entities (sometimes via merging/splitting of unsupervised clusters) through manual approaches that require prior biological knowledge and curation based on known cell markers, or through reference mapping and label transfer from annotated atlases [76]. Clusters are thus discrete groups of cells (be it types or states), so to attempt and reconstruct the structure of the continuous biological process being studied, trajectory inference methods such as Slingshot [97] and PAGA [98] have been developed. These trajectories are mapped onto an inferred pseudo-time axis, and this process can be further complemented by RNA velocity. RNA velocity [99] is a method that refers to the usage of splicing kinetics to model transcriptional dynamics and infer vectors of tanscriptional change (i.e. the direction and rate of gene expression change) along the manifold of cells. These vectors can either be used on their own to infer a pseudo-time axis [100], or act as an input layer for further downstream analyses that attempt to determine cell fates (as opposed to or in addition to cell states and types) [101].

Compositional analysis refers to the methods used to explain how structure in the cell space is affected by perturbations under different conditions. The first single-cell *omic* studies presented relatively simple experimental designs (due to high costs and low throughput), and thus their work tended towards the description of a particular condition. However, as technologies have advanced and costs have trended downward, more complex experimental designs have emerged where it is pivotal to model and quantify the effects of perturbations (e.g. mutations and drug treatments in the context of CRC). Hence the emergence of compositional analysis methods in recent times, such as Differential Abundance [102, 103], MELD [104], and TrajectoryNet [105]. Furthermore, there are also a set of approaches to *in silico* model perturbations that were not part of the experimental design [106, 107, 108], but these methods tend to struggle when modelling genes with low expression values.

While trajectory analysis and RNA velocity are extremely useful for determining cellular dynamics in a differentiation setting, a cell-based metric of pluripotency is

also of special interest to discern stem cells from differentiated cell fates. To this end the concept of Signalling Entropy Rate was postulated [109], which argues that the higher the entropy of a cell's transcriptomic profile, the less differentiated and thus higher pluripotency degree it presents. Currently there are several methods to estimate cell pluripotency from scRNA-seq data, most relying on signalling entropy rates and computationally faster approximations like the degree of correlation between the transcriptome and Protein-Protein interaction matrices [110, 111, 112].

Exploring the structure within the feature space is key towards understanding the biology at a mechanistic (and not just descriptive) level. In the context of scRNA-seq, structure in the feature space is commonly determined through differential gene expression (DE), which determines the degree and statistical significance of changes in a gene's expression across individual cells or groups of them (e.g conditions, labelled cell identities or cellular neighbourhoods). The most common DE methods are pseudo-bulk approaches derived from the mature field of bulk sequencing [113, 114] or population comparison tests like the Wilcoxon signed-rank test. These methods are commonly applied to compare clusters, in which case they generate a list of markers characteristic of each cluster/population, but might also be used to compare conditions or even cellular neighbourhoods [115]. The resulting gene markers can then be passed through Gene Ontology [116] and pathway databases [117, 7, 118], or Gene Set Enrichment Analysis tools [119] to identify putative biological processes for each cell group.

Much like in the context of cell structure, k-NN graphs of genes can also be constructed from either interaction databases or gene expression data. These graphs can then be used to determine gene modules and gene regulatory networks [120], and represent a relatively unexplored avenue for emerging methods when compared to the much more common cell-graphs. Cell-to-cell communication tools also leverage these interaction databases with the aim of inferring cellular interactions through the co-expression of ligands, receptors, and other interaction member genes [121]. Methods like CellPhoneDB [50] and CellChat [49] predict ligand-receptor interactions by identifying clusters of cells that express receiving or

sending members of the interactions, and can be used together with spatial studies to refine their predictions [122, 123]. Given the broad diversity in methods for determining an interaction and the different interaction databases used, ensemble methods such as LIANA have been designed to aggregate often conflicting cell-cell communication results [124, 125].

1.4.4 Limitations and New Avenues

Accessibility and scalability advancements to single-cell multiomic technologies are empowering a complex and multifactorial view of cell identity. This is especially relevant in the field of cancer research, where our understanding is shifting from the canonical genotype-driven cancer cell state toward plasticity-driven phenotypes.

However, this nuanced view of cell identity clashes with the concept of cluster derived cell types, especially those derived from transcriptomic data that could be argued are better suited to capture a cell's state. Furthermore, our understanding of biological processes wherein cells represent individual points along a continuum is not really suited to discrete cluster-based groups. In response to this necessity, there has been a series of emerging cluster-free approaches, such as the concept of cellular neighbourhoods as applied by John Marioni's lab, or the notion of cellular archetypes and metacells. The cellular neighbourhood approach was first implemented as miloDA in the context of compositional analysis [103], and has recently been adapted for DE tasks [115]. They iterate on the concept of clusters defined on a k-NN graph to that of cellular neighbourhoods; which both contain fewer cells than a typical cluster and can overlap over the same regions of the graph. Cellular archetypes and metacells represent a more orthogonal way of tackling the limitations of cells clusters, as rather than aiming to capture discrete cell types they aim to capture cell states [126, 127]. Thus, within each metacell state, all cells should ideally represent the same biological state defined by a unique profile of gene regulatory programmes and only be distinguished by technical noise. With new methods developed to address multiomic data and cross-patient integration [128], metacell-based approaches appear perhaps poised to replace the ubiquitous unsupervised clustering approaches. This view of cells as landmarks on a continuous landscape is far from a

novel concept. In the mid 20th century, Conrad H. Waddington illustrated the process of an epigenetic landscape where pluripotent cells would roll down into valleys of terminally differentiated states [129]. However, his effort and subsequent ones since then have mostly been of a rather subjective and artistic nature. Reconstructing such landscapes from biological data is not an untenable task anymore, as omic profiles from single-cells can be embedded together and mapped onto a 2D space. Sculpted by cellular pluripotency metrics, such landscapes have already been proposed, but used embedding spaces that do not accurately reconstruct a continuous space that captures global structure and did not leverage information on transcriptional dynamics [130].

The idea of cell-cell graphs derived from gene or protein data is also central and common to virtually all single-cell omic analyses, including scRNA-seq and mass cytometry. k-NN graphs of feature nodes however are a less exploited niche, often relegated to the study of gene regulatory networks and systems biology approaches. However constructing such graphs is not a trivial task, for coexpression metrics generally do not capture gene-gene interactions, most gene regulatory networks do not account for directionality [131], and curated interaction databases [7] are not consistently analysed in a directed way. Hence I explore a novel approach of assembling directed gene-gene knowledge graph (KG)s and then projecting cells into the graph based on their transcriptional profile, thus treating the cells as signals on a gene graph. Similar methods with comparable goals are emerging [132], suggesting a neeed for further method development in this field.

1.5 Hypothesis and Aims

Organoids represent a robust model able to recapitulate CRC dynamics and its interaction with the TME. The high dimensional information captured by single-cell *omic* approaches and the diverse field of analyses promise the potential of untangling and describing even the most complex of biological processes. In light of this, **I hypothesise that colon-epithelia polarisation by endogenous and exogenous cues can be described using single-cell analyses of organoids**.

First I present my efforts identifying and solving gaps in the method space that can facilitate mass cytometry analyses broadly. In Chapter 3 I introduce CyGNAL, a workflow that aims to facilitate standard MC data analysis steps for a non-computational audience. Additionally I also discuss and showcase the use of machine learning approaches to automate cell-state classification for MC data.

To test the main hypothesis I aim to perform a comprehensive and state-of-the-art single-cell analysis of CRC organoids to: 1) systematically describe the colon epithelial stem regulation, and 2) *in silico* infer mechanisms of regulation that have been subsequently tested *in vitro* by colleagues [1]. Chapter 4 presents the main corpus of results from this analysis. In Chapter 5 I present a novel method to generate data-driven Waddington-like landscapes that capture the underlying continuous processes of transition and differentiation, and I demonstrate how they can be used to model the landscape of colon epithelial stem regulation.

Finally in Chapter 6, I further my aim towards solving a lack of methods for both intra- and inter-cellular communication analyses by exploring a KG-based approach to study cell communication in organoid-fibroblasts co-cultures. Appendix A presents *pyKrack* a standalone tool and package for computing hierarchy scores on directed graphs, such as a cell-communication interaction graphs.

Chapter 2

Materials and Methods

2.1 CyGNAL

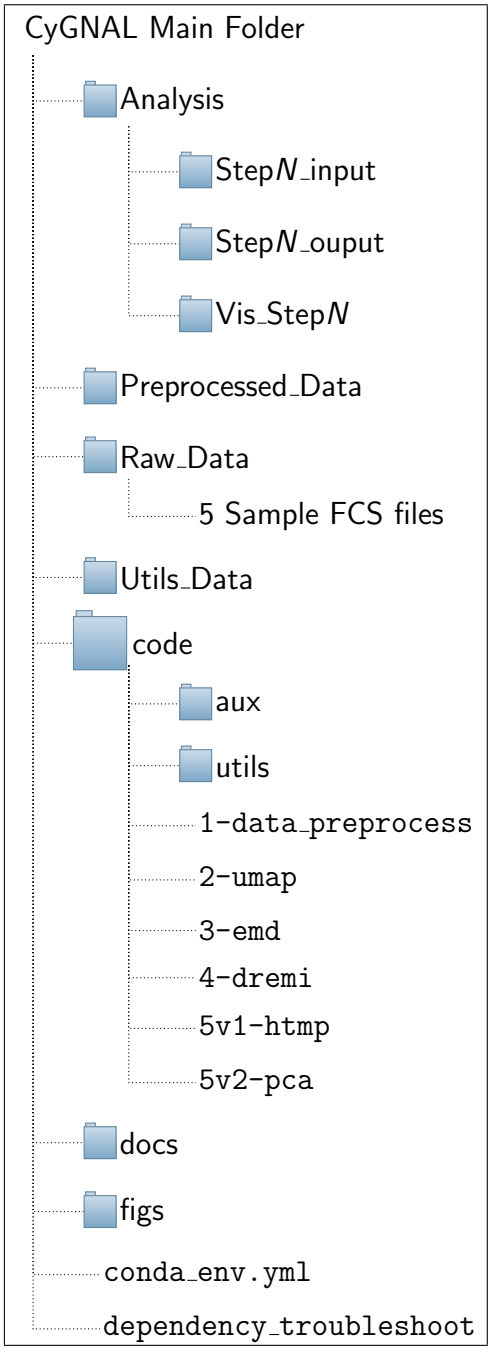


Figure 2.1: CyGNAL Directory Tree. Main steps are in the code directory, with Input/Output achieved through the data and analysis folders.

Written mainly in Python and R, CyGNAL (CyTOF SiGNalling AnaLysis) is a pipeline constituted as a series of core scripts within the `code` directory. Considered as the main steps of CyGNAL, these scripts have been numbered according to the canonical order within CyGNAL's workflow. The first script handles data pre-processing and must always be run. The second script embeds cells in a two-dimensional UMAP space. The third and fourth steps compute the EMD and DREMI scores, which can then be visualised using either Heatmaps in step `5v1` or as through PCA in step `5v2`.

Python modules containing function definitions are kept within the `aux` directory, while the `utils` directory contains optional steps and utility scripts for data handling. The resulting modular structure allows for general utility functions used throughout CyGNAL to be defined once within a single file. Data ingestion and egestion is done through a series of input and output directories, either specific to each of the main steps or common for all scripts within the utilities folder.

2.1.1 Deployment and Dependencies

CyGNAL is intended to be deployed as a non-standardised pipeline by cloning the directory to a local machine. However, to minimise any possible dependency issues, CyGNAL includes a Conda environment `YML` file. Conda (`anaconda.org`) is a package and environment management system that works with multiple programming languages, including Python and R. Hence, with the included `YML` file, a software environment with all of CyGNAL's R and Python dependencies might be replicated in a single step.

However, there are cases when the Conda environment fails to solve and a suitable environment can not be generated, such as when using different compute architecture. For these instances I have also prepared a containerised distribution method using Docker [133]. Based on an x86 Debian Linux container, CyGNAL's container automates the process of creating a Conda environment with all required dependencies on platforms with a different architecture like ARM-based Apple Sillicon.

This container is hosted on Docker Hub and can be *pulled* from `docker.io/ferranc96/cygnal:one`.

Running CyGNAL from the Docker container only requires of two additional steps:

- Pull CyGNAL from the GitHub repository, place it in your home directory (i.e. `~`), and rename the *CyGNAL* folder to *CyGNAL_docker*.
- Run the following command on the host terminal:
  ```
  docker run -v ~/CyGNAL_docker/:/usr/app/CyGNAL -it
  --entrypoint /bin/bash -p 12241-12252:12241-12252
  docker.io/ferranc96/cygnal:one.
  ```
- Use CyGNAL commands on the container terminal as by running the individual python scripts in the `code` directory.

The docker command above runs a live terminal on the container with a Conda environment that already contains all necessary dependencies. Communication with

Function	Main Regex	Description
rename_columns	(__[a-z].*$\| __\d.*$| _\(.*$\| ___.*$)	Expression catches badly formatted channel names with doubled or tripled underscores, so that the function can simplify channel names.
filter_columns	^\d+[A-Za-z]+$	Pattern that matches a string that starts with one or more digits followed by one or more letters until the end of the line.

Table 2.1: Pre-Processing Regular Expressions. Column renaming and filtering is achieved via regular expressions for fuzzy text and pattern matching.

the host machine is done via the shared directory in /CyGNAL_docker (i.e. where you will need to input data and fetch CyGNAL's outputs), with open ports to access the Heatmap and PCA shinyApps. It is important to note that on first run it will take some time to pull the image (1.5GB), and that alternative container tools such as Podman (podman.io) should work but are not officially supported.

Furthermore, should the user encounter any issues while using CyGNAL, the Python script dependency_troubleshoot.py should help locate and report to the user any missing dependencies.

2.1.2 Computation

During the pre-processing step CyGNAL loads in mass cytometry files either as tab-separated plain text format or in the Flow Cytometry Standard (FCS) format (FCS)[134]. Intercompatibility between both formats is ensured using the Python packages *fcsparser*[135] and *fcswrite*[136], and the R package *flowCore*[137]. In addition to ensuring format consistency and allowing for datasets to be saved in either format, during the pre-processing step channel names are parsed to; a) eliminate empty channels; b) clean up double spaces and underscores; and c) ensure each cell has a unique ID encoded in a new column called "Cell_Index". This is accomplished using the rename_columns and filter_columns functions via regular expressions (Table 2.1).

Finally, this first pre-processing step also writes to disk a panel_markers.csv file containing those columns present in the dataset that were identified as markers (i.e.

where the channel name is composed of an isotope and an antibody or other cellular marker. The panel_markers.csv file can then be used by the user to filter out certain channels for downstream steps.

CyGNAL's Universal Manifold Approximation and Projection (UMAP) calculation uses the *umap-learn* package [91] to embed the cells in a 2-dimensional space. The embedding is computed using the set of markers defined by the user in the panel_markers.csv file and can be calculated on either just one processed dataset or a series of datasets as long as they have shared markers in their panel. The resulting coordinates are appended as a new pair of columns to the original datasets, facilitating visualisation of this space elsewhere by the user.

EMD stands for Earth Mover's Distance and is named so because it can be intuitively thought of as the amount of work required to transform between two piles of earth, where work refers to the mass of earth to be moved times the distance. Also known as the 1st Wasserstein distance (W_1), it is defined between two 1D arrays of measured values u and v as:

$$W_1(u,v) = \int_{-\infty}^{+\infty} |U - V|$$

Where U and V are the cumulative distribution functions of u and v respectively.

Applied to the mass cytometry datasets in CyGNAL, I score each marker (chosen via the panel_markers.csv file) based on its distribution of intensities in a variable dataset (u) when compared to a particular reference (v, either defined from the sum of all datasets imputed or a particular dataset selected by the user). The absolute value of the distance metric is then signed based on the median values of the variable and reference distributions in order to assign a direction to the changes observed that can then be interpreted in a biological setting (e.g we want know how much the apoptotic marker cCaspase 3 [D175] changes between a condition and the control, but also where its median intensities are higher).

Described in Van Djik *et al.* [138], k-NN conditional Density Resampled Estimate of Mutual Information (DREMI) is a mutual information metric that reflects how informative the distribution of intensities for marker A is in describing the

intensities of maker B (i.e. $I(A|B)$). Unlike the EMD scores that compares across conditions then, DREMI is computed on a per condition basis, where each of the possible combinations of markers in panel_markers.csv is scored.

Both the EMD and DREMI scores are computed using the Python package *scprep* [139], and the outputs of both scoring systems are saved as plain text files that can be plotted using CyGNAL's visualisation steps below.

It is important to note that for calculating the EMD and DREMI scores and computing the UMAP space, the data is by default normalised using an inverse hyperbolic *asinhx* transform with a co-factor of 5. However, the user is prompted to override the default behaviour if so desired, and the co-factor value can be easily changed within the various scripts.

2.1.3 Visualisation

CyGNAL automates and allows for the user to visualise both EMD and DREMI scores in an interactive manner via Shinny-Apps [140].

The Shiny-Apps are contained within R files loaded from the last main scripts of the CyGNAL workflow. For this, user defined arguments in the python scripts need to be parsed to the R Shiny server when it is called through bash using Rscript.

The first of the visualisation scripts generates a series of heatmaps using the *ggplot* [141] and *ComplexHeatmap* [142] packages. These heatmaps show the relevant scores; with the names of the datasets used in the calculation step as columns in the horizontal axis and the names of the markers in the vertical axis as rows. Colour ranges, columns, and rows shown can all be tweaked by the user through the graphical interface. The second of the scripts computes a PCA on the scores using the *FactoMineR* package [143], treating each of the datasets used in the calculation as observations and the scores for the markers (or marker pairs in the case of DREMI) as variables. In all cases, all plots generated can be saved as images for later use, and within the PCA Shiny-App the computed PCA coordinates can also be downloaded to facilitate custom generation of plots elsewhere.

2.2 Cell-State Random Forest Classifier

2.2.1 Design and Architecture

The cell-state classifier built uses the *scikit-learn* Python package [144] to train a Random Forest classification algorithm and assign cell-state labels to mass cytometry datasets. A Random Forest (RF) algorithm is based on a series of decision trees, simple non-parametric models that predict the class of an observation by learning decision rules inferred from the data during training. By using a randomised collection of these trees (i.e., a forest) the RF palliates the tendency of decision trees towards overfitting while at the same time reducing the variance of the results. This is so because each of the individual trees sees only a subset of the data, hence they built different models. Then, being an ensemble method, when each datapoint is passed through them all, a majority vote decides on the class given.

Hosted in `https://github.com/FerranC96/C_StateML`, this cell-state RF classifier consists of two Python scripts and shared auxiliary functions. The first of the scripts is used to train a model from labelled data and report on its performance against validation and testing datasets. Default parameters are used for the Random Forest (except for an increase in the number of decision trees to 480), and the cell-state classes in the training data are balanced by donwsampling to the least common state. Pre-trained models are also included in the repository as will be detailed below. Balancing classes is done to ensure that all cell-state classes are trained using the same number of cells and so that performance metrics such as F_1 scores, which are vulnerable to imbalanced classes, can be used.

The second script is used to run a saved RF model through new mass cytometry datasets to label and assign a state to each cell. While designed to work with unlabelled data, if the input data is already labelled this script also reports on the model's performance.

Performance evaluation is reported both as text and in the form of plots, and consists of; 1) confusion matrices, 2) log losses, 3) precision, recall, and F_1 scores for each class.

For class c, let:

$$precision_c = \frac{TP_c}{PredP_c}$$

where precision is also known as Positive Predictive Value, TP_c is the number of true positives, and $PredP_c$ the number of predicted members in c,

and let

$$recall_c = \frac{TP_c}{P_c}$$

where recall is also know as True Positive Rate and P_c is the number of cells in c,

the F_1 scores for each class c are defined as the harmonic mean of the precision and recall of c so that:

$$F_1 = 2\frac{precision \cdot recall}{precision + recall}$$

Before both training and evaluation the data is assumed to be in the form of raw intensities and gets transformed using an inverse hyperbolic *asinhx* transform with a co-factor of 5.

2.2.2 RF Classifier Models

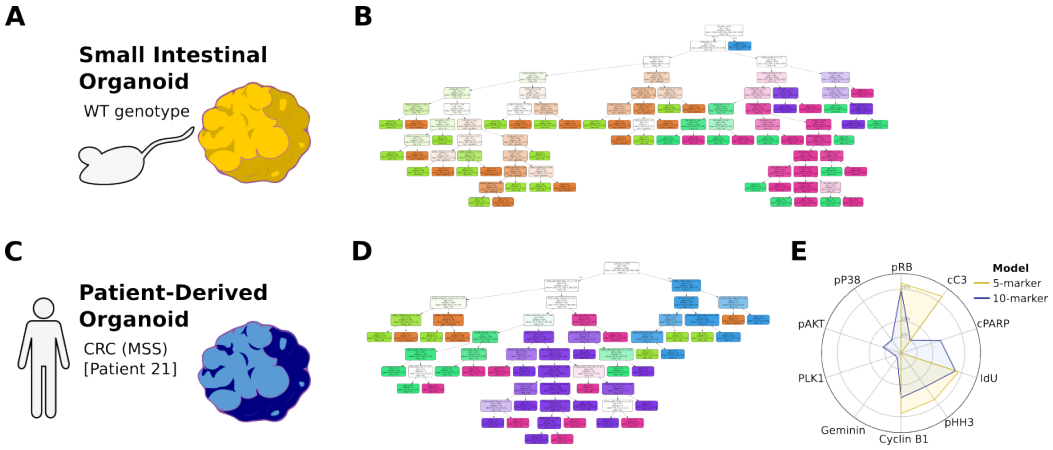


Figure 2.2: Training Random-Forest Cell-State Classifiers. A) 5-marker model trained with intestinal organoids and, **B)** structure of a single tree. **C)** 10-marker model trained with intestinal organoids and, **D)** structure of a single tree. **E)** Comparison of feature importance between the models. MSS, micro-satellite stable.

This classifier implementation was used to build two models distinguished by the features they use and the type of epithelial cells they were trained on.

Marker	Specificity	Model
pRB [S807, S811]	Proliferating cells	5-marker, 10-marker
pAKT [T308]	PTM in the mTOR pathway (proliferation)	10-marker
pP38 [T180, Y182]	PTM promoting β-cat activation (proliferation)	10-marker
pHH3 [S28]	M-phase	5-marker, 10-marker
PLK1	G2/M-phase transition	10-marker
Cyclin B1	G2	5-marker, 10-marker
IdU	S-phase	5-marker, 10-marker
Geminin	Negative marker of G1. Expressed in S-phase, G2, and M-phase	10-marker
cCaspase 3 [D175]	Apoptosis	5-marker, 10-marker
cPARP [D214]	Apoptosis	10-marker

Table 2.2: Markers Used in the RF Models. Antibody markers and their targets used in the two cell-state classifier models.

The simpler 5-marker model (Figure 2.2A) uses only 5 cell-state markers and was trained using a balanced subset of cells from the Small Intestinal murine organoid time-course experiment in Qin *et al.* [4]. This model uses the same exact antibody markers as those used by Qin *et al.* to label cell-state via manual gates, namely: pRB [S807/S811], cleaved Caspase 3 [D175], IdU, Cyclin B1, and pHH3 [S28].

The more complex 10-marker model was trained using CRC Patient Derived Organoids (Figure 2.2C). With an updated panel, the markers used in the latest models are a set of ten antibodies (the five markers from above plus cPARP [D214], pAKT [T308], pP38 [T180/Y182], Geminin, and PLK1) with targets specific to each of the six cell-state classes (Apoptosis, G0, G1, S-phase, G2, and M-phase). The data used to train this model has been published in Ramos Zapatero & Tong *et al.* [3] and belongs to an untreated monoculture replicate of PDO21.

Details on the markers used in the RF models, and the cell-state they are associated with, can be found in Table 2.2.

2.3 scRNA-seq Data Analysis

Work presented in this section has already been made public in Qin & Cardoso Rodriguez *et al.* [1] (Appendix D). As a joint co-first authored paper, attribution is shared between Dr. Xiao Qin and myself. While I carried out all of the scRNA-seq Data Analysis presented in this Thesis, Dr. Xiao Qin was in charge of the murine colonic organoid culture system and data acquisition via both scRNA-seq and Mass Cytometry. The exact attribution for specific tasks is detailed in Qin & Cardoso Rodriguez *et al.*

Aiming to provide additional context, the section below on scRNA-seq data acquisition has been included despite Dr. Xiao Qin having carried-out the work. For details on the organoid platform used and the general experimental setup see Qin & Cardoso Rodriguez *et al.* [1].

2.3.1 Data Acquisition

In brief, the organoid heterocellular culture system was dissociated into single-cells, FAC-Sorted for live cells, counted and fixed with methanol before scRNA-seq library preparation. For co-cultures, different cell-types were mixed at equal cell numbers prior to the fixation step. scRNA-seq libraries were generated with the 10X Genomics Chromium Next GEM Single Cell 3' Reagent Kits v3.1 (Dual Index) and sequenced with the Illumina NovaSeq 6000 System (2$\times$ 150 bp paired-end reads), aiming at 60,000 read pairs per cell and 2,000 cells per cell-type per sample. For more details see Qin & Cardoso Rodriguez *et al.* [1].

2.3.2 Data Processing

The Illumina NovaSeq binary base call (BCL) outuput sequence files were converted to FASTQ files and processed with the 10X Genomics Cell Ranger pipeline version 5.0.1 [145], which provides with a convenient wrapper for Illumina's bcl2fastq tool [146], `cellranger mkfastq`. Prior to alignment, a custom murine GRCm38-based reference genome was generated using the STAR aligner [79] wrapper `cellranger mkref`. By adding the sequences for *DsRed* and *eGFP* transgenes present in fibroblasts and organoids respectively, cell-type discrimination based on exogenous

transcripts was facilitated. Then, alignment of the FASTQ files against this custom reference was performed using the `cellranger count` pipeline, generating both unfiltered and pre-filtered feature-barcode matrices.

The resulting feature-barcode matrices were analysed with the R package *Seurat* version 4.0.4 [147]. The analysis pipeline encompasses quality control, data normalisation, data integration, dimensionality reduction, cell clustering, and analysis of differential gene expression. Genes found in less than 4 cells were removed during QC and only cells with at least 600 unique genes identified were kept for downstream analysis. The total number of detected reads per cell typically ranged from 1,200 to 80,000, with the actual values manually determined based on dataset sequencing depth and cell-type composition. Cell-type composition was considered as the macrophages were observed to be captured less efficiently than fibroblast or epithelial cells. For the integrated epithelial object in used throughout the analysis, an additional filtering step was performed to remove cells with undetectable expression for any one of the *bona fide* pan-epithelial genes *Epcam*, *Krt8*, *Krt18*, *Krt19*, *Cldn7*. Doublet/multiplet filtering was explored using the scDblFinder package [148], which has been designed to find heterotypic doublets such as those that could be present in co-culture conditions. However, after the QC pipeline outlined above, the low number and homogeneous distribution on 2D embeddings of the putative doublets was not deemed convincing enough to warrant their removal.

The Seurat object, much like its SCE counterparts in R or AnnData objects in Python, contains multiple layers where different $barcode X feature$ matrices can be stored. This is so it can accommodate for different normalisation methods and for tools that expect raw or processed gene expression values.

Gene expression values were normalised for total counts, multiplied by a scale factor of $1 \cdot 10^4$, and the expression values log transformed as $X = \log_2(X+1)$. The resulting normalised count matrix was used for methods that rely on the explicit comparison or visualisation of gene expression.

An alternative normalisation approach was used as described in Hafemeister & Satija 2019 [85]. Named sctransform (SCT) this method models both biological and

technical variation using the Pearson residuals from a regularised negative bionamial regression. The nature of this model allows for certain signals to be regressed out, such as the percentage of mitochondrial transcripts over the total reads in a cell, or for differences between cycling cells in different phase of the cell cycle. I computed SCT normalised count matrices with 6,000 features and regressed out mithoncondrial content and differences between cell cycle phases.

Cell Cycle scores were computed using the CellCycleScoring function from Seurat (a wrapper for AddModuleScore) and a curated list of cell cycle genes shown in C.3. By comparing how well a cell matches the G2 and M-phase signature, versus a G1 and S-phase signature, cells could be classed into Dividing cells (Mitosis and G2), Cycling (G1 and Synthesis), and Other (cells with low scores for both signatures, most likely outside of the cell cycle). Differences between the Dividing and Cycling groups were also regressed out when computing the SCT normalised count expression data, as it was deemed that intra-cycle differences were not central to the biological system being studied. This computation of cell cycle scores used a custom table of cell cycle genes shown in Sup. Table C.3.

Throughout the study, steps that relied on building a k-NN graph or low-dimensional representation of the data use either the SCT normalised data or the SCT-derived integrated representation (see section below).

2.3.3 Integration

Dataset integration was performed using Seurat's reciprocal PCA (RPCA) implementation [147] as it has been optimised to handle large datasets. RPCA works by projecting the individual datasets into an other's PCA space to identify cellular anchors with shared neighbourhoods across projections. The integration itself is described in Stuart *et al.* [149], so that new expression matrices in the integrated space are computed based on the difference of expression matrices between anchor cells. Inherently a pairwise process, integration of multiple datasets is done iteratively by pairs accordings to their pairwise distances. In this work I used the SCT normalised data as the feature space to be integrated and ran default parameters but for a k.anchor of 12. The integrated object presented in Figure 4.1 was computed

using all cells from the 20 conditions shown in the figure, resulting in a total of 58,726 cells with the integrated assay limited to 2,000 genes. The integrated object first presented in Figure 4.2 and found elsewhere across this work was computed using just the epithelial cells from all conditions, resulting in an object with 29,452 cells limited to 4,000 genes. The respective integrated feature spaces were stored within the integrated assay of the Seurat object.

The integration pipeline with anchors found via Canonical Correlation Analysis was also tested, but as described in the literature, it was found to be less computationally efficient and appeared to smooth out and erase too much biological signal [149].

When handling the aggregated data from multiple CRC patient cohorts presented in Joanito et *al.* [6], where data integration was performed mostly for visualisation purposes, the methods described above struggled to handle the high number of cells present (>78,000 cells including projected organoid data). As one of the goals of this data integration approach was to compare our murine organoids with human samples, I used scVI [150], a Variational AutoEncoder approach that can be GPU-accelerated and performs well on inter-species integration tasks [151]. Part of a broader family of PyTorch-based methods for analysing single-cell *omic* data [152], scVI learns a low dimensional latent space that can be used to compute 2-dimensional embeddings of the data. Able to account for multiple quantitative and categorical confounding variables, this method can also handle the projection of query datasets onto an integrated reference. Cross-species data integration was thus achieved by generating an integrated reference from the human CRC datasets (filtered to the top 6,000 most variable genes) and projecting into it a humanised version of the mouse organoid data from Figure 4.2. The integrated human reference was built using unique patient identifiers as the batch key and controlling for the percentage of mitochondrial reads in a cell. The resulting latent space was embedded into 30 dimensions. Humanisation of the mouse count matrix was accomplished via the mousipy package [153], which facilitates the handling of mouse genes with multiple human orthologues. Untransformed count data was used for the scVI workflow.

2.3.4 Dimensionality Reduction

To generate the EMD PCA plots shown in Figure 4.1C I used the normalised gene expression data of all cells of a particular cell-type (organoids, fibroblasts, or macrophages) stored within the RNA assay of the integrated Seurat object from Figure 4.1B. EMD scores for the top 6,000 variable genes of each condition were computed with CyGNAL [154] using the relevant control condition for each cell-type: WT monoculture for epithelial organoids, fibroblast monoculture for fibroblast cells, and macrophage monoculture for macrophage cells. The collection of gene-specific EMD scores for each condition was then used to compute a PCA space where each dot represent a whole condition.

The standard pipeline for generating single-cell embeddings consisted of computing a set of 50 to 100 principal components (PC) from a normalised count matrix, from which 2-dimensional PHATE embeddings were generated with default parameters. PHATE was chosen as the default DR method for visualisation due to its capacity to capture the global structure in biological settings with important developmental trajectories [92]. In the context of integrated datasets via scVI, the 30-dimensional latent space was used to generate the PHATE embeddings. This mid-dimensional PCA space was also used to compute most of the k-NN cell-cell graphs used throughout the study.

2.3.5 Unsupervised Clustering and Differential Expression

Cell clustering was computed using the Leiden algorithm on the k-NN graph generated from the integrated epithelial dataset (first 48 PCs), at a series of resolutions ranging from 0.2 to 0.8. The final cluster annotations were retrospectively defined by curated cell-type marker expression (Figure 4.2C), inter-cluster relationships on a multi-resolution clustering tree [155], and cross-condition differential abundance behaviours (Figure 4.3). Cells from outlier clusters (totalling less than 1% of all epithelial cells) were excluded from the downstream analysis (Figure 4.2A).

Differentially Expressed (DE) genes between clusters, conditions, and cell neighbourhoods were identified using Wilcoxon rank-sum tests as implemented in Seurat's *FindAllMarkers* and *FindMarkers* functions. The Wilcoxon rank-sum test is

commonly used in the field of scRNA-seq as a non-parametric test, albeit with the assumption that the samples compared are independent. DE results are presented in the form of log transformed fold changes in gene expression, with p-values adjusted for multiplicity of tests.

Heatmaps of selected marker genes were generated with the R package *ComplexHeatmap* [156]. Gene lists in Figures 4.4 were curated from previously reported markers for colonic epithelial subpopulations and DE genes detected between epithelial clusters, conditions, and DA neighbourhoods within this study. Gene lists in Figures B.2, B.1, and B.3 represent DE genes between conditions.

2.3.6 Differential Abundance

Differentially abundant (DA) cell neighbourhoods were identified using the R package *MiloR* [103]. Milo works by constructing cellular neighbourhoods on a k-NN graph. These neighbourhoods can overlap with one another, for cells may belong to multiple neighbourhoods at once, and act as the basis of Milo's compositional analysis. By comparing the composition of these neighbourhoods in terms of a categorical variable of interest (condition), Milo assigns them an enrichment score (log Fold Change) according to the relative abundace of cells from the query or control condition. Significance and regression out of technical and unwanted biological variables is achieved though a Generalised Linear Model via the mature edgeR package [113], and using the SpatialFDR metric (first described in Lun *et al.* [102]). DA analysis thus allows for the detection of enrichment and depletion of epithelial cell states caused by microenvironmental and/or genotypical perturbations in the organoid system.

For the analysis shown in Figure 4.3A-B I set the DA test threshold at 5% SpatialFDR. In the context of fibroblast regulation of the colonic epithelia, given that $CD34^{hi}$ and $CD34^{lo}$ fibroblasts do not differentially regulate epithelial cells (Figure B.2), all samples of WT organoid+fibroblast co-cultures were grouped and considered replicates of the query condition regardless of the CD34 status of the fibroblasts. AK and AKP organoid monocultures were also grouped due to their similar DE and DA behaviour (Figures 4.4, 4.3C). The DA overview dot plot in Figure

4.3C was generated by comparing the 17 conditions against the WT monoculture control (2× replicates). Absence of replicates in this approach results in a lack of relevance for the SpatialFDR statistic, and the control condition (1st row) was populated with empty values for visualisation purposes.

The *k*-NN graphs used by Milo where constructed as detailed in the section above.

2.3.7 Signature Score Correlations

By gathering more than 50 gene lists from the literature that describe key signalling pathways and stem-related gut epithelia states [30, 157, 32, 158, 159, 160, 161, 162, 163, 164, 165, 166, 167, 168, 169, 28, 29, 170], I could put the results of this study in context with a broader corpus of works in varied settings; such as human data, cancer, or tissue repair processes. Gene lists for different intestinal stem cell-states were compiled from public datasets, together with transcriptional targets of key signalling pathways associated with the different stem cell-states. Gene identifiers where transformed to murine Gene Symbols (to be compared with the features in the sequencing dataset) by querying BioMart [171]. Metadata for the resulting compiled list can be found in Sup. Table C.9.

These gene lists were compared with the curated gene signatures for proliferation, CSC, revCSC, and proCSC cell-states in Figure 4.4, as well as the top DE genes for each stem cluster (adjusted p-value <0.01, log2FC >0.25, top 24 genes with the greatest positive log2FC values). *UCell* scores for each gene set were calculated using log-normalised gene expression values and z-scored to allow cross-signature comparison.

The *UCell* [5] method was used to generate the correlation matrix between gene signatures in existing literature and cell clusters identified within this study (Figure 4.7A). UCell uses the Wilcoxon rank-sum test, also known as Mann-Whitney U, and a matrix of ranked genes (by expression) for each cell in the dataset. Given a list of genes, UCell can then score how well each cell matches their expression. Unlike Seurat's *AddModuleScore* function (and its re-implementaion in scanpy), the resulting scores are not normalised against a control gene set, making UCell scores

impervious to dataset composition.

Pearson correlations were computed between the relevant scores (i.e. those whose context refers to the stem compartment or key signalling pathways) on all cells of stem and TA clusters and then visualised as a heatmap-like correlation matrix with the corrplot package [172]. Signatures that did not have a SD deviation of scores greater than 0.2 across the cells where excluded from the analysis (as they would equally mark all states present in the dataset), and matrix entries where populated for significant correlations (confidence level of 0.95). Finally, the matrix was ordered and grouped via complete linkage hierarchical clustering (k=3).

2.3.8 Signalling Entropy and Pluripotency

Leveraging the concept that cells with a higher potency should have a higher signalling entropy [109], the pluripotency values for epithelial cells across the different clusters were estimated using the R package *SCENT* [110]. Signalling entropy scores for all epithelial cells in Figure 4.5A were determined via the CCAT approximation method, which computes a Pearson correlation between a cell's transcriptome and the interactome as defined by the built-in *net17Jan16* Protein-Protein interaction network (derived from the Pathway Commons database). As the interaction network is annotated with NCBI gene IDs, BioMart was used to translate them to MGI gene symbols.

Being a method that is completely independent of any cell metadata, like clusters or conditions, the resulting vector of CCAT scores was added as a new metadata column to the sequencing dataset object and used to quantify pluripotency changes in Figure 4.5B and as one of the components of the Valley-Ridge score (Chapter 5).

2.3.9 RNA Velocity and Cellular Dynamics

For RNA velocity analysis, loom files were generated from Cell Ranger's output using the command line interface tool *velocyto* [99]. The murine GRCm38 reference was used, with the GRCm38/mm10 repeat mask assembly and the RepeatMasker track. RNA velocity was analysed with the Python package *scVelo* [100] using close to default parameters. Metadata and PHATE embedding coordinates were exported

from the relevant Seurat objects to filter and annotate AnnData objects generated from the loom files made by velocyto. Moments for the velocity estimation were calculated using the first 50 PCs and 30 neighbours from the AnnData objects. RNA velocities were computed with the *recover_dynamics* function using the dynamical model of transcriptional dynamics with default parameters. The velocity stream embedding (Figure 4.5A) was computed using the integrated object containing epithelial cells from all conditions. The RNA velocity vector lengths, an estimate of a cell's rate of transcriptional change, were computed using cells solely from the 4 conditions shown in Figure 4.5C. The quantitative comparison in Figure 4.5D was performed using the Games-Howell pairwise test wrapper from the R package *statsExpressions* [173]. All conditions were compared against the WT monoculture control and all p-values have been corrected for multiplicity with the Holm method.

Initial and terminal macrostates where determined using CellRank [101], which leverages RNA velocity information to describe cellular dynamics. The matrix of cell-cell transition probabilities was constructed as a weighted combination of the transition matrix based on velocity directions (through the VelocityKernel class, weight of 0.8) and a symmetric transcriptional similarity matrix (through the ConnectivityKernel class, weight of 0.2). Macrostates, and their transition probabilities, were computed using the built-in Generalised Perron Cluster-Cluster Analysis (GPCA) estimator. To find initial macrostates, inverse velocity vectors are used to assemble the transition matrix by setting the backward argument to `True` when computing the VelocityKernel component. Directed PAGA plots [98, 101] were computed so that epithelial clusters are represented as nodes shown on top of a low dimensional embedding and are connected by directed edges whose thickness represents local velocity flows.

2.3.10 Cell-Cell Communication Analysis

Cell-cell communication inference was performed using the R package CellChat [49], where stromal-epithelial signalling was analysed across 4 different organoid genotypes (WT, A, AK, and AKP). CellChat uses a database of interactions between ligands, receptors, and cofactors. Using cluster annotations and the gene expression

matrix, interaction probabilities can be inferred between the different populations using a permutation-based approach. The inferred interactions can the be grouped at the pathway level, and functional analysis of the clusters can be infered via network analysis methods.

Epithelial cells were annotated with the clusters previously identified (Figure 4.2A), while the fibroblasts were grouped as a single cluster. A merged CellChat object was generated to compare relative communication probability of fibroblast-to-epithelia signalling across the genotypes. Significant ligand-receptor pairs were identified based on CellChat's murine cell communication database. Plots displaying aggregate outgoing and incoming communication probability (Figure 4.6A) were generated with the *netAnalysis_signalingRole_scatter* function. Detected communication at the pathway and interaction level was accessed with the *subsetCommunication* function and probabilities were z-score normalised to allow for cross-pathway or cross-interaction comparison. The results were visualised with ComplexHeatmap in Figure 4.6B, the rows of which were manually ordered based on hierarchical clustering and grouped based on the nature of the interaction. Gene expression of the ligand-receptor pairs identified above was visualised using Seurat's *Dotplot* function in Figure 4.6C. *UCell* scores for ligand and receptor genes were calculated for fibroblasts and epithelial cells respectively and quantified in Figure 4.6D. Games-Howell pairwise test was performed using the R package *statsExpressions* and all p-values have been corrected for multiplicity with the Holm method.

2.4 VR Score and Data-Driven Waddington-like Landscapes

Work presented in this section has already been made public in Qin & Cardoso Rodriguez *et al.* [1]. I am the author behind the Valley-Ridge (VR) score design and implementation, including the python-based renders of the data-driven Waddington-like landscapes.

Dr. Jeroen Claus however, kindly rendered the landscapes shown in Figure 6 of Qin & Cardoso Rodriguez *et al.* using the professional rendering software *SideFX: Houdini*. The exact attribution for specific tasks is detailed in the manuscript [1].

2.4.1 VR Score Computation

The VR score is cell-based metric defined as the weighted sum of the Valley and the Ridge components (Figure 5.1):

$$VR = 0.9V + 0.1R$$

where V is the Valley component and R the Ridge component.

The Valley component is computed as

$$V = med(CCAT)_{s,c}$$

for each combination of sample (s) and cluster (c).

Let u be scaled representation of the velocity vector length for each cell ($\frac{1}{|v|}$), and d be the scaled median L^1 distance of each cell to all other cells from the same cluster. d acts a cell centrality metric computed on a k-NN graph of a cluster PHATE embedding, followed by the calculation of a shortest distance matrix (using the graphtool software [174]) whereby cells with the lowest median distance would be at a cluster's centre whilst those with the highest distance would be at the cluster periphery. Outliers with a distance over Q_{99} were set to the median distance. To allow for inter-cluster comparisons, d was scaled for each cluster to the (0,1) range with sklearn's *MinMaxScaler* [144], whereas u was scaled at a dataset level using

the same function. The Ridge component is then computed per each cell as

$$R = med(u)_{s,c} \cdot d$$

This definition of the VR score allows the CCAT-based Valley component to be the driving force for sculpting the landscape and the velocity-driven Ridge component to predominately define local features at the boundaries between clusters, producing a tarn-like effect symbolising a state of trapped cells in cluster whose cells present low velocity vector lengths. In principle, any other dimensionality reduction technique can be used in place of PHATE [130], and the Valley/Ridge component can be computed using other metrics underpinning pluripotency and cell-fate transition. The Ridge component can also be calculated with a distance-free approach such as α-shapes [175]. Finally, the VR scores could be computed on a per cell or neighbourhood basis, which would increase landscape resolution and liberate the method from constraints of cluster definitions (at the expense of increased noise).

2.4.2 VR Landscape Projection

To generate the Waddington-like landscapes in Figure 5.1B, I combine the ability of PHATE to capture the global structure of single-cell data with the VR score.

Waddington-like landscapes can be visualised directly in Python (Figure 2.3. Briefly, a low dimensional 34x30 mesh grid was generated from the PHATE embeddings, and a 3D surface was rendered by projecting VR scores onto the grid using the radial basis function interpolation from scipy [176]. The surface of the landscape was coloured by VR scores and a scatter plot was overlaid where the elevation of each cell was defined as the weighted sum of its VR score (weight = 0.9), CCAT value (weight = 0.1), and a constant factor of 0.012 (weight = 1). This added a level of controlled noise to the scatter plot while ensuring most cells remain above the interpolated surface.

Finally, external software can also be used to render the data-driven landscapes, as shown in Qin & Cardoso Rodriguez *et al.* where we used the 3D rendering programmes SideFX Houdini and Maxon Redshift (Figure 2.3).

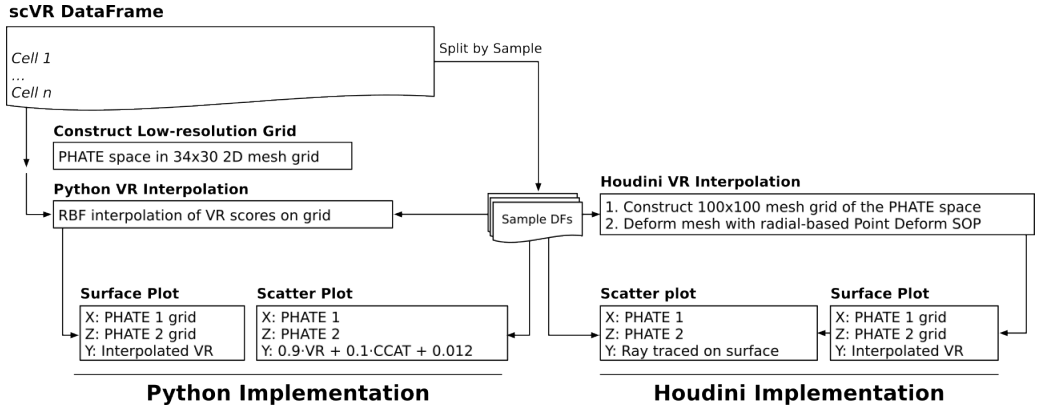


Figure 2.3: Generation of Waddington-like Landscapes from scRNA-seq Data. Workflow for landscape projection using the built-in python-based method or the alternative approach with Houdini. VR, valley-ridge.

2.5 Knowledge Graphs for Cell Communications

2.5.1 Sources and Assembly

Interaction information found in cell communication databases can be parsed and formatted as a knowledge graph (KG). To assemble the custom Ligand-Receptor-Target KG (LRT-KG) I accessed the CellChat [177] and NicheNet [178] databases. Ligands and receptors were gathered from both databases, whereas transcription-factor (TF) target genes were extracted from NicheNet's. Formatted as lists of human gene symbols, the three categories (ligands, receptors, and targets) were pruned to ensure there was no overlap between them, simplifying the KG and enhancing its hierarchical nature.

The KG was assembled as a table of relational triplet entries wherein a *head* node interacts with a *tail* node via a *relation* edge. This relational information was obtained from the Reactome database of curated pathways [118]. Assembly of the KG was thus achieved by iterating through all possible *head* and *head* combinations and creating a *relation* between them if both were found to belong to the same pathway level in Reactome's second level of pathway hierarchies.

The *de novo* assembled custom LRT-KG was compared with the popular curated repository of cellular interaction knowledge OmniPath [7], which contains almost four times the amount of nodes present in the LRT-KG but with a similar number of relations, resulting in a slightly lowered average degree (Table 2.3). By processing

Graph	Nodes	Edges	Pathways	$\overline{\text{Degree}}$	Hierarchy
LRT	2507	97054	23	77.43	1
Omnipath	9248	92262	*NA*	19.95	0.82
Omnipath (processed)	9248	94836	33	20.51	0.78

Table 2.3: Knowledge Graph Characteristics. Table comparing KG metrics between the LRT-KG and the Omnipath repository. LRT (KG), ligand-receptor-target (KG).

the OmniPath database as detailed above and incorporating the pathway information from Reactome we observe how the number of distinct pathways present is also higher than in the LRT-KG, with 33 and 23 unique pathways respectively (Table 2.3).

Hierarchy of the assembled graphs was computed using my python package *pykrack* (`pypi.org/project/pykrack/`), which computes the Krackhardt hierarchy score for a given directed graph (see Appendix A for more details). The OmniPath graph is highly hierarchical before and after processing, and the LRT-KG hierarchical design results in a completely hierarchical tree-like structure (Table 2.3).

2.5.2 Embedding the Knowledge Graph

The table of relational triplets was then used to generate a KG using the MultiDiGraph() function from the NetworkX package [179], wherein multiple types of edges (*relations*) connect nodes in a directed manner.

The resulting directed KG was then embedded into a lower 50-dimensional space using the TransR KG embedding algorithm [180] as implemented in the PyKEEN package [181]. TransR is a knowledge graph embedding approach derived from the mature TransE algorithm [182] that better encodes complex relational information.

The embedding space was learnt on GPU compute using an 80:10:10 train, test, and validation split and otherwise default parameters. To visualise the embedding the 50-dimensional space was embedded using PHATE (Figure 6.3), or projected as a graph whose layout was determined via the Fruchterman-Reingold force-directed

algorithm as implemented in NetworkX [183]. Metadata was then added to the node-based embedding based on the presence of the gene symbols as ligands, receptors or TF targets in the cell communication databases discussed above (Figure 6.3A). Furthermore, pathway-level metadata from Reactome was also used to check for the presence of genes in each of the Reactome pathways (Figure 6.3B).

2.5.3 Wavelet Transform and Data Projection

Through data projection I can evaluate *omic* profiles of cells as signals on a k-NN graph derived from the KG embedding. First the k-NN graph is computed using the sklearn package [144] (n_neighbors = 5). Then, on each of the nodes of the graph a wavelet bank is centred at and diffused at J scales, resulting on a flattened $nodeXwavelets$ matrix; wherein nodes equal the number of genes in the KG, and the wavelet bank equals genes times the scale parameter J (with all data shown using $J = 4$). A high-level overview of this process is presented in Figure 2.4, where a single diffusion wavelet is shown centred around a particular node of the Stanford bunny graph [184] (Figure 2.4A). A bank of wavelets at four scales ($J = 4$) is computed for the different nodes of the graph (Figure 2.4B).

This wavelet computation is based on a python script kindly provided by Aarthi Venkat from Prof. Smita Krishnaswamy's lab at Yale University (Table C.2), which implements the wavelet definition from Coifman & Maggioni [89] wherein a diffusion wavelet transform is the difference between two scales of lazy diffusion on a graph. github

To generate the projected $cellXwavelets$ matrix I compute the dot product between the $nodeXwavelets$ matrix and the $cellXfeature$ count matrix with the numpy package [185] (Figure 6.4B). The shared feature axis between the two matrices is aranged by filtering and reordering the features in the gene count matrix to match the nodes of the KG.

Resulting in a relatively high dimensional space (with cells on one axis and the product of features and the wavelet scale J parameter on the other), PCA is applied to the projected matrix and then a k-NN graph and subsequent PHATE embedding are computed. The PHATE embedding serves as a useful non-quantitative way to com-

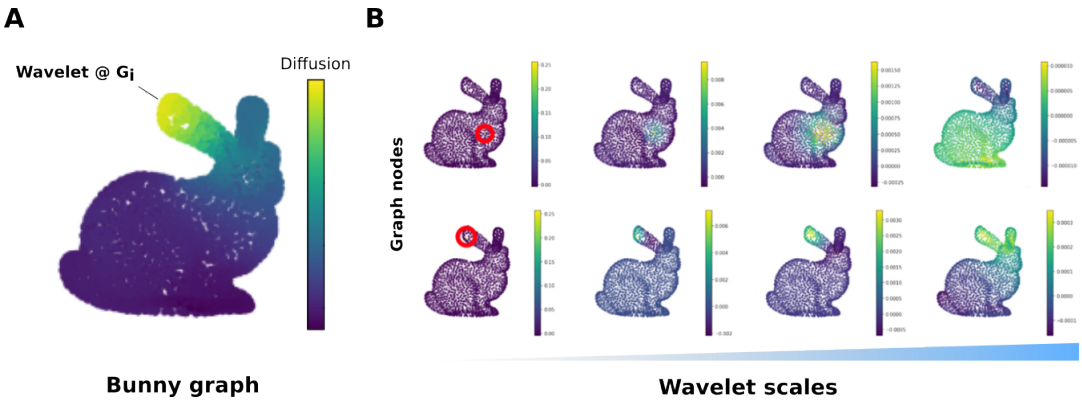


Figure 2.4: Diffusion on KGs with a Bank of Wavelet Transforms. A) Wavelet transform is centred on a graph's node to diffuse a signal. **B)** Bank of wavelets at multiple scales is applied to each node in the graph.

pare the projected data with the original count matrix-based embedding (Figure 6.4). The projected k-NN graph can be used to quantitatively compare and benchmark the method against a k-NN computed directly from the $cellX feature$ count matrix (Figure 6.5. Distances on the k-NN graphs were computed using the shortest_path() function from the graphtools package (`github.com/KrishnaswamyLab/graphtools`) and then aggregated at the cluster level by computing the mean distances from and to each pair of cell clusters (Figure 6.5). To compare these inter-cluster distances between the gene expression and LRT-KG projection spaces (Tables C.11 and C.12), the distance matrices were scaled to (0,1) range using the MinMaxScaler() function from sklearn and substracted to generate a matrix of distance differences (Figure 6.5B). Pearson correlation between the two unscaled distances for each cluster-pair combination was computed with the scipy package (Figure 6.5C), and so was the correlation between the distances and the aggregate communication probability score for each cluster pair as defined by CellChat (Figure 6.5D).

2.6 FAIR Spirit and Reproduceability

Furthering the spirit of shared scientific knowledge and collaborative research embodied by the FAIR principles, data and code used to generate the analyses in Qin & Cardoso Rodriguez *et al.* have been made public in various repositories. Furthermore, tools and outputs developed during my project and presented in this thesis have also been made publicly available aiming to make my research FAIR: Findable, Accessible, Interoperable and Reusable.

These outputs have been disseminated either as part of peer-reviewed publications such as CyGNAL in Sufi & Qin *et al.* [2], as software packages like pykrack (Appendix A, `ferranc96.github.io/pyKrack`), in the form of publicly accessible GitHub repositories (such as `github.com/TAPE-Lab/Qin-CardosoRodriguez-et-al`), or even as entries in my personal blog (e.g. `ferranc96.github.io/posts/GSPw23`).

For details on the software and tools used to write this thesis and make its figures see Appendix F. The source code and original figure and data tables used to generate this thesis are currently part of a private repository that will be made public once the final version has been approved and entered UCL's registry (`github.com/FerranC96/FerranCardoso_ThesisPhD`).

Chapter 3

Building Accessible and Automated Mass Cytometry Analysis Tools

3.1 Introduction

As outlined in Chapter 1, the Thiol Organoid Barcoding *in situ* (TOB*is*) mass cytometry (MC) platform used to analyse the colorectal cancer (CRC) organoids is already a mature approach. The effects of both tumour microenvironment (TME) and genotypical pertubartions in this organoid system were already explored [4], but data analysis was performed using custom and discrete scripts; encumbering consistency and reproducibility for future analyses. Furthermore, the manual process of cell-state annotation added further load to the analysis.

To improve upon this I have designed and developed CyGNAL (CyTOF SiGNalling AnaLysis) [154], a pipeline for MC data analysis with a focus on studying post-translational modification (PTM) changes across multiple conditions. CyGNAL aims to streamline and bring to non-computational scientists analyses similar to those shown in Qin *et al.* [4], with the addition of dimensionality reduction embeddings and interactive visualisations. CyGNAL was published as part of Sufi & Qin *et al.* [2] in conjunction with an updated TOB*is* custom mass cytometry platform for organoids (Appendix E).

The maturity of the platform is also reflected on the properties of the markers used in the MC panels, with the most robust markers achieving highly binary and specific staining. Given the importance of cell state changes to perturbations in the epithelial organoids, either in the form of intrinsic effects such as genotype or extrinsic in the form of the TME or drug treatments, an automated approach of labelling and assigning a cell state to each cell in an experiment would facilitate routine analysis of MC datasets. I thus hypothesise that we can use a machine learning approach to, using a series of canonical cell state markers, automatically predict and label the hundreds of thousands of cells captured in an MC experiment. To this end I aim to develop a Random Forest (RF) classifier. This classifier will be able to ingest MC data and, using manually gated datasets with cell state labels as training data, label each of the cells with one of six possible cell states: Apoptosis, G0, G1, S-phase, G2, and M-phase.

3.2 CyGNAL: CyTOF Signalling Analysis pipeline

Published and demoed as part of Sufi & Qin *et al.*, CyGNAL is a publicly available tool that is routinely used to analyse MC datasets both at my group and by external collaborators [186]. Details on the implementation, code structure and deployment can be found in Chapter 2. Furthermore, a step-by-step walk-through of the main CyGNAL steps is detailed in Sufi & Qin *et al.* [2].

In this section I will present an overview of the tool and will discuss the relevance of the different scoring systems with regards to MC data in general and PTM signalling panels in specific. Example outputs from CyGNAL will also be shown; both for the computational sections (scores and UMAP embedding), and how they can be further analysed, but also with screenshots of the interactive apps that constitute CyGNAL's visualisation steps.

3.2.1 Overview and Capabilities

CyGNAL is a collection of scripts written mainly in Python and R. These scripts have been built around a unified code base of shared functions and a particular directory structure to facilitate interoperability between the different steps. Within CyGNAL's code directory, the `utils` folder has optional steps that either complement the main ones or contain additional utilities for MC data handling.

Distribution of CyGNAL is accomplished as a container hosted in Docker Hub (hub.docker.com/repository/docker/ferranc96/cygnal). CyGNAL can also be used by downloading the project's public repository (from github.com/TAPE-Lab/CyGNAL) and then installing all required Python and R dependencies via conda using the provided YML environment file. More details on this process can be found in Chapter 2.

The tool relies on the computation of two scores, Earth Mover's Distance (EMD) and Density Resampled Estimate of Mutual Information (DREMI), to analyse the intensity of detected antibodies across conditions or other gating-derived metadata groups (i.e. cell-cycle phase or cell type). EMD (also known as the Wasserstein distance) is an optimal transport metric that describes the distance between distributions of detected intensities, and thus is used to compare protein/PTM expression across

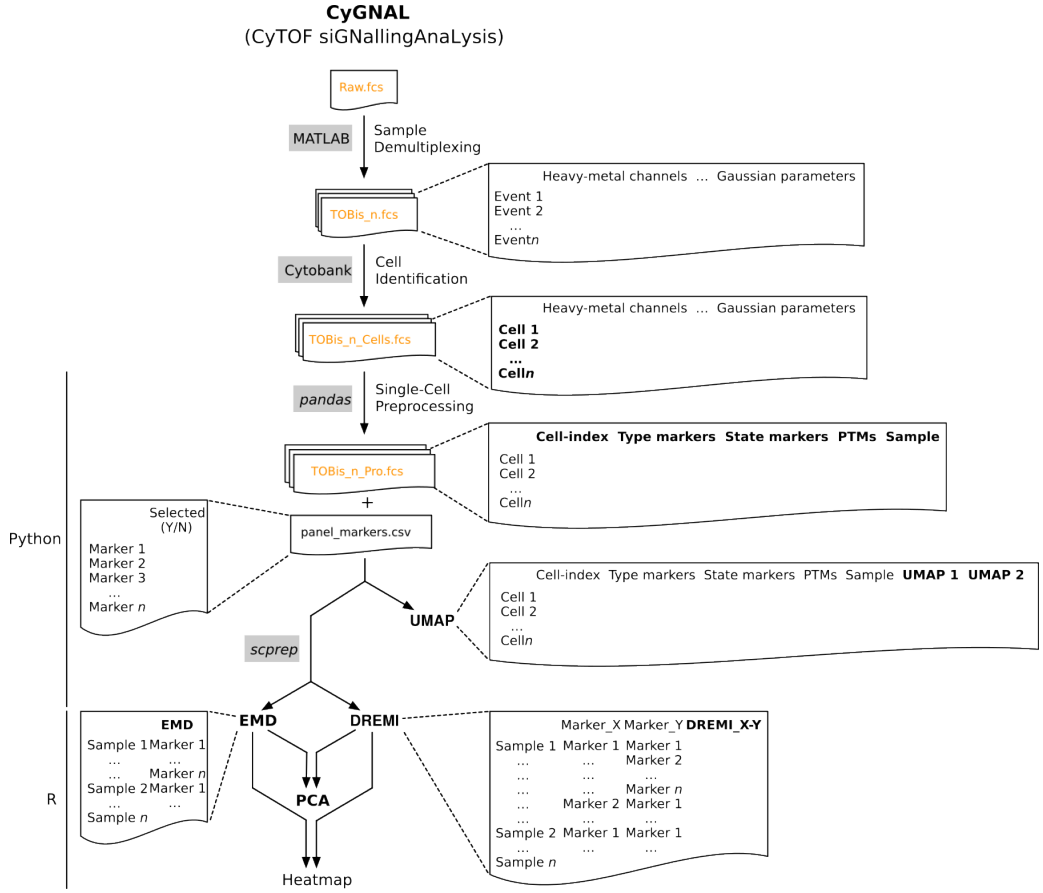


Figure 3.1: CyGNAL MC Analysis. File structure and processing architecture of CyGNAL in the context of upstream processing steps. Software environments and packages are indicated in grey boxes, computational processes are in bold text. DREMI, Density Resampled Estimate of Mutual Information. EMD, Earth Mover's Distance. PCA, Principal Component Analysis. UMAP, Uniform Manifold Approximation and Projection.

experimental conditions. DREMI is a mutual information estimate that can be used to relate the degree of connectivity across conditions of protein/PTM pairs. More details on both methods and how they are implemented can be found in Chapter 2.

A general overview of CyGNAL's structure is shown in Figure 3.1, where the tool encompasses the bottom two thirds of the diagram.

As CyGNAL uses FCS files or tab-separated plain text files, certain upstream processes are necessary after data acquisition. Previous to analysing the data in CyGNAL, the standard operating procedure in our lab is to debarcode the mass cytometry datasets in MATLAB (using the tool from `github.com/zunderlab/single-cell-debarcoder`) and perform initial data pre-processing and quality

control in Cytobank (`www.cytobank.org`). In that platform, the single cells are gated for Gaussian parameters, their DNA content, and uptake of Cisplatin using manual gates. Gating on cell-state and cell-type specific markers can also be done in order to both eliminate doublets but also to identify cells belonging to each state or type; information which can then be used to understand the biological system, but also train the cell-state classifier.

The CyGNAL workflow starts with a pre-processing step. Here, empty heavy metal channels with no conjugated antibodies are removed, and the remaining channels are renamed to reduce the presence of special characters and keep with the naming conventions of the Fluidigm CyTOF software. A unique cell identifier is also given to each cell, and experimental metadata can also be embedded within the main pandas dataframe. Furthermore, a file with updated antibody channel names is also saved (panel_markers.csv), so that the user can select which channels to use in downstream steps.

Dimensionality reduction via Uniform Manifold Approximation and Projection (UMAP) [91] can be performed to embed the individual cells on a 2-dimensional space based on the selected antibodies.

EMD and DREMI scores are computed using the scprep package [139]. Compute time can be reduced by subsetting the panel to channels of interest, and the user gets prompted to define specific arguments relevant to either computation, such as defining the variable and reference distributions for the EMD step.

Finally, the computed EMD and DREMI scores can be visualised as heatmaps or further summarised via PCA to compare profiles across conditions using CyGNAL's last two main steps. The visualisation steps load in the default and user-given parameters and pass them to R Shiny-Apps [140] that host a local server which automatically opens on the browser.

3.2.2 Use Case and Outputs

CyGNAL is distributed with sample mass cytometry datasets, which originate from technical replicates of an organoid monoculture experiment. They have been down-sampled so that they can be hosted on GitHub and distributed with the code itself.

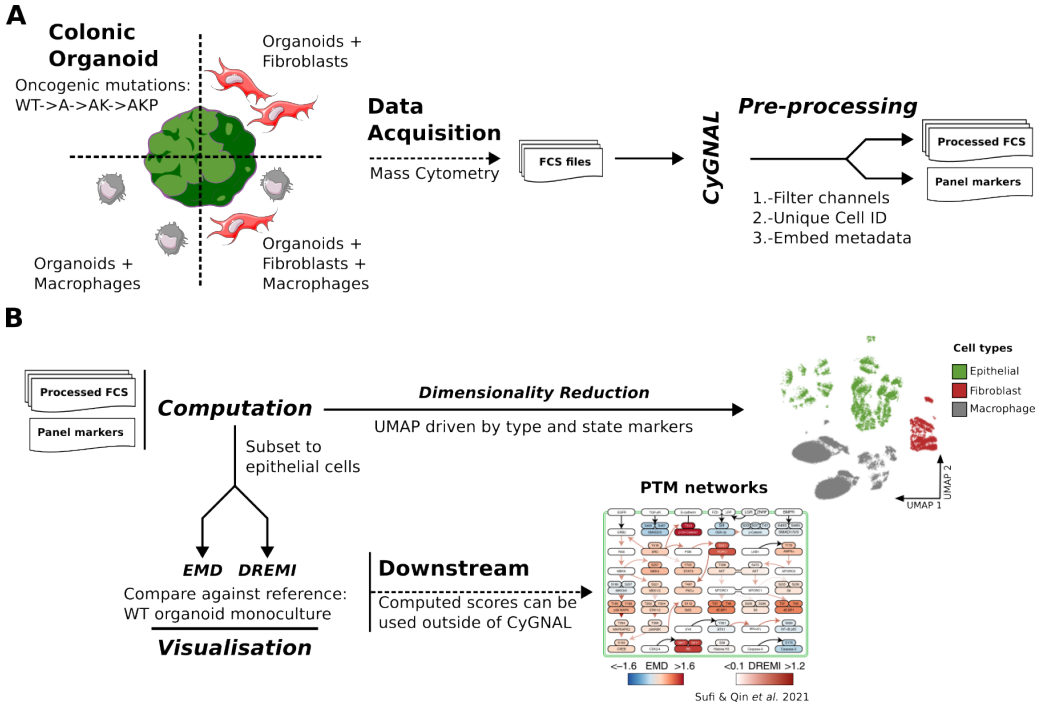


Figure 3.2: Analysis of Heterocellular Organoid Systems with CyGNAL. A) Pre-processing steps follow MC data acquisition. **B)** Computation steps generate low-dimensional embeddings and scores that can be visualised downstream. WT, wild-type. A, *shApc*. AK, *shApc* and $Kras^{G12D/+}$. AKP, *shApc*, $Kras^{G12D/+}$ and $Trp53^{R172H/-}$.

The results presented in Figure 3.3 A-C were generated with this sample data.

In Figure 3.2 I present a mass cytometry dataset from Qin *et al.* [4] to showcase an example use with heterotypic culture conditions where cell-type-specific analysis is necessary. The data belongs to the same mouse colon organoid model from Chapter 4 and presents with a similar experimental setup, wherein organoids with different genotypes where cultured on their own or with macrophages and/or fibroblast cells. Data was subsequently gated and annotated on cell types and states as described above and on the original publication [4], and then passed onto CyGNAL for pre-processing (Figure 3.2A).

Cell state and type markers were then selected using the panel marker file (pHH3, IdU, cCasp3, pRB, LRIG1, CEACAM1, pan-CK, F4/80, PDPN, RFP, CyclinB1, CD68) to generate a UMAP embedding (Figure 3.2B). This low-dimensional embedding resolves the three distinct cell-types (Figure 3.2B).

Using the cell-type gates previously drawn on Cytobank, unique cell identifiers

were used to select only the organoid cells. Computation of EMD and DREMI scores was then performed on the epithelial compartment, and can be visualised as part of CyGNAL. Furthermore, in the specific context of PTM network signalling analysis, EMD and DREMI scores can be used to assemble signalling network diagrams. With signed EMD used to quantify PTM node intensity and DREMI to score PTM-PTM edge connectivity, a signalling network can be curated and manually annotated as shown in Qin *et al.* [4]. When paired with a well-curated antibody panel and robust experimental design, TOB*is* MC allows multiplexed analysis of cell-type–specific PTM signalling of heterocellular organoids [2].

Using the sample data and with the concatenation of all input files as the reference for the EMD step, Figure 3.3A demonstrates CyGNAL's heatmap visualisation. By selecting not to use a specific reference during the EMD computation step, the generated scores are useful to compare how antibody expression compares across each of the individual datasets/conditions. The heatmap ShinyApp lets the user control the colour scale (automatically set to maximise contrast on the range of EMD scores), remove antibodies from the heatmap, and reorder the datasets/conditions shown in the columns. The heatmap shown in Figure 3.3A is an interactive version generated with Plotly [187], and shows the corresponding EMD score when hovering over a cell. Furthermore, a similar non-interactive heatmap is generated using the ComplexHeatmap [142] package and can be found within its homonymous Shiny-App tab.

The same data was used when running the PCA Shiny-App in Figure 3.3B-C. This CyGNAL steps lets the user explore the data by looking at the raw scores (Figure 3.3B) and Pearson correlation between channels. The user can also define parameters for the Principal Components Analysis, including the number of markers, generate several types of PCA plots with or without eigenvectors overlaid, and export the PCA results as plain text. In Figure 3.3D I demonstrate how, despite CyGNAL being originally designed to handle mass cytometry data, other types of single-cell omic data like scRNA-seq can also be used. Here I used CyGNAL to compute EMD scores based on the gene expression of the organoids sequenced in Chapter 4 and generate a

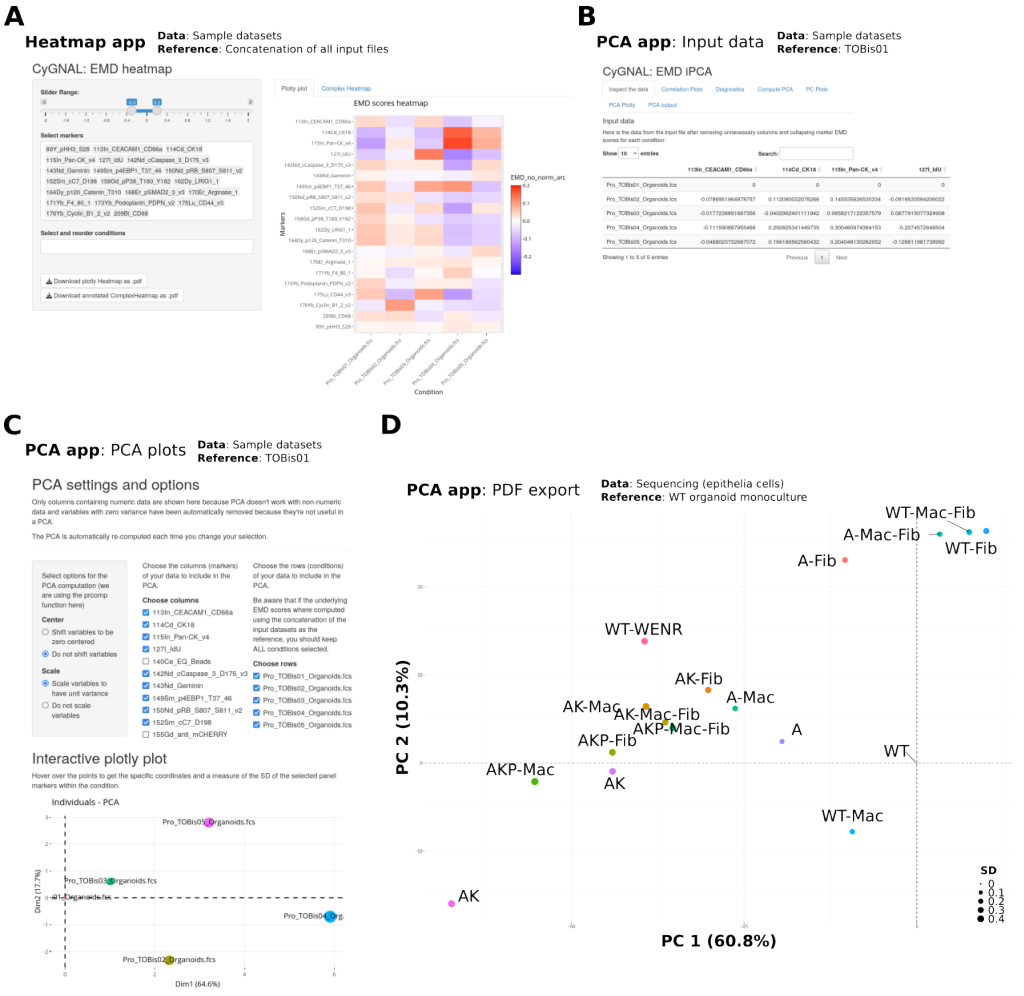


Figure 3.3: CyGNAL Outputs Interactive Visualisations. **A)** Screenshot from CyGNAL's interactive heatmaps, using computed EMD scores from the example data provided. **B-C)** Screenshots from CyGNAL's interactive PCA computation and visualisation, using computed EMD scores from the example data provided. **D)** Editable vector graphics PCA plot generated using EMD scores from scRNA-seq data.

PCA embedding showing how the different conditions compared to the control. Note that the PCA data in Figure 3.3B-D was generated using EMD scores computed with a particular dataset/condition as the reference and without centring the PCA embedding matrix. This application serves as an example of use-cases where there is a clear control condition against which the other conditions are compared to (like the WT organoid monoculture in Figure 3.3D).

3.3 Cell-State Random Forest Classifier

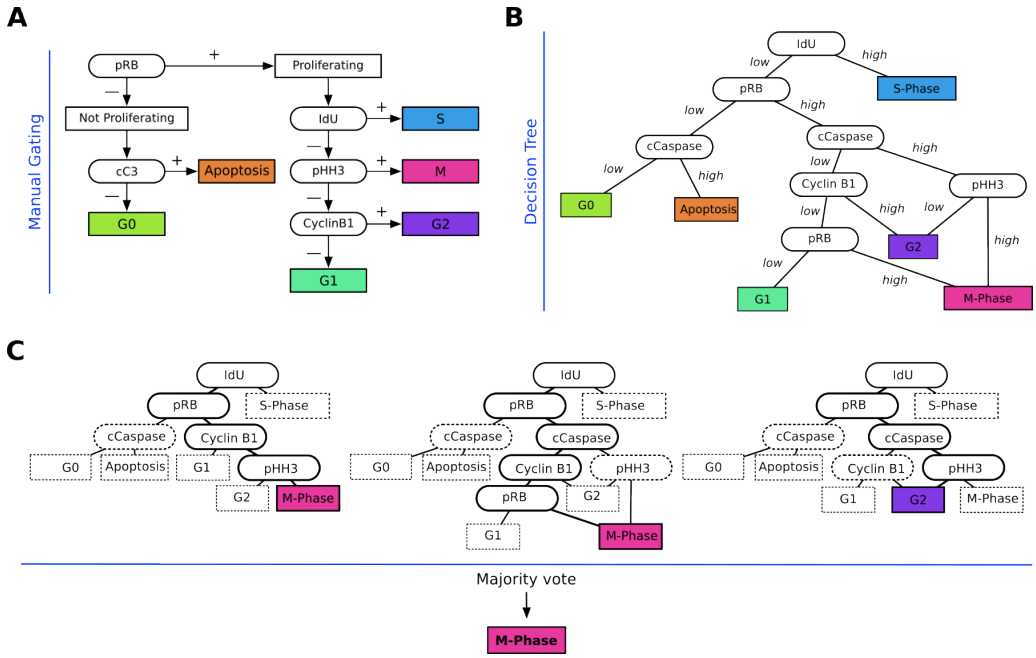


Figure 3.4: Forests of Decision Trees that Resemble Manual Cell-State Gating. A) Process of manual cell-state gating from Qin *et al.* [4]. **B)** Schematic representation of a cell-state decision tree with binary logic gates. **C)** Random forests as an ensemble of decision trees.

Determining a cell's state with regards to the cell-cycle phases is central to understanding the intestinal epithelium response to perturbations, as shown by Qin et *al.* and their observations regarding cell-type specific regulation of cellular states in response to microenvironmental and oncogenic cues [4].

The cell-state labels are commonly established using manual gating on a biaxial marker state, wherein a researches draws a boundary separating 2 groups of cells, essentially thresholding the data based on antibody expression [4] (Figure 3.4A). However, generating these cell-state labels is a time consuming process, especially when compounded with the scalability of MC and TOB*is* ability to perform highly multiplexed analyses. Issues with user-induced biases are also present, as drawing the manual gates is a subjective process that might not remain consistent from experiment to experiment.

Early on my PhD I was exploring the link between PTMs and cell-state when I noticed that the process of generating the cell-state labels could potentially be

automated using a classical supervised machine learning approach. Eventually, I developed a cell-state Random Forest (RF) classifier to automate this process (see Chapter 2 for more details). The manual gating process naturally resembles the logic behind a decision tree, as in both a threshold of antibody intensity would result in a binary classification of cell groups (Figure 3.4A-B). Furthermore, the RF machine learning approach remains a white box whose internal decision logic can be easily interpreted, for it consists of a collection of individual decision trees trained on subsets of the data that are used together in an ensemble approach (Figure 3.4).

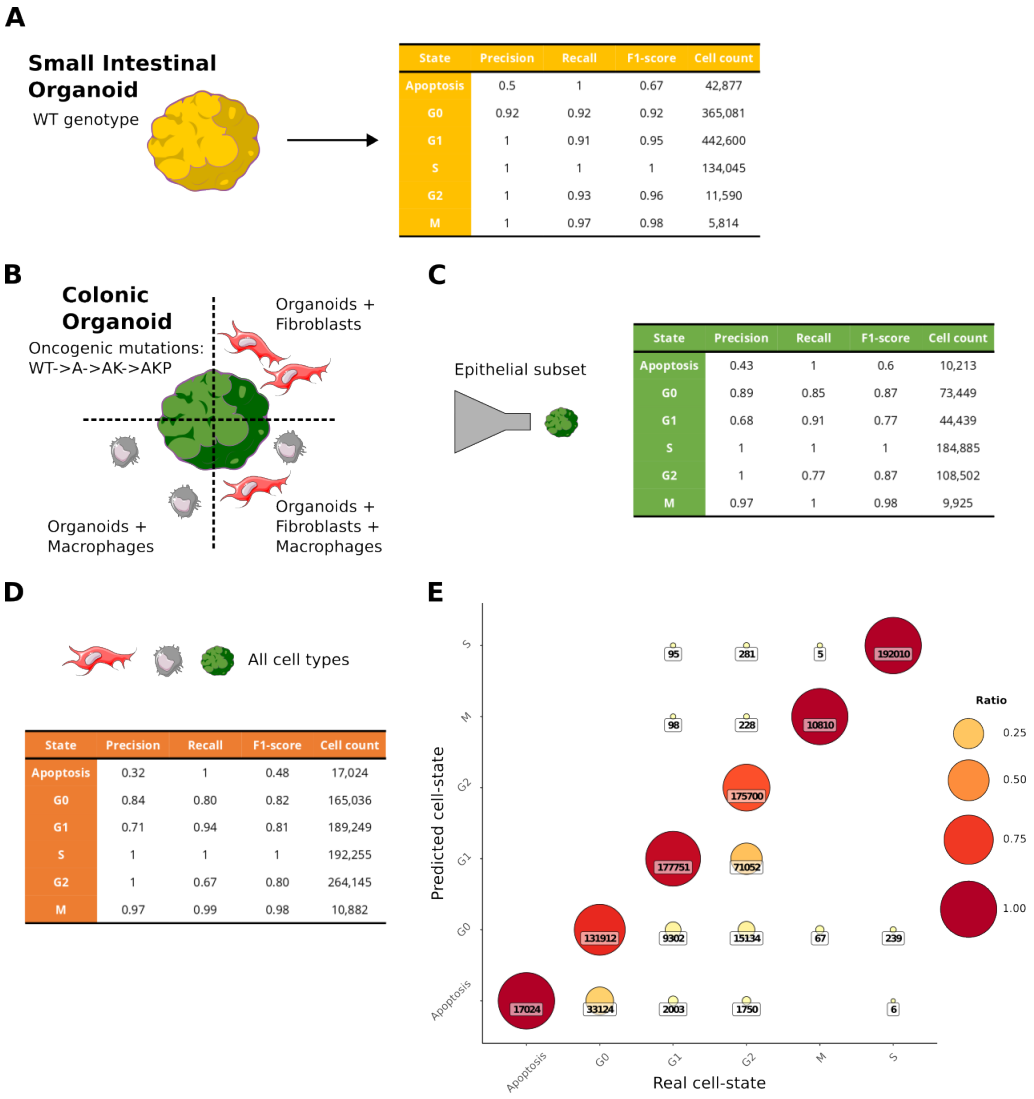
3.3.1 5-marker Model Performs Across Model Systems

The first RF model built was trained on data from the murine small intestinal organoid cultures from Qin *et al.* [4], consisting of WT organoids along several developmental time-points (Figure 3.5A). This model used only the 5 markers shown in Figure 3.4A. Details on building the model and the relative feature importance when training can be found in Chapter 2.

Testing the 5-marker RF model on a different single time-point small intestinal organoid dataset also from Qin *et al.* results in global accuracy for all classes of 0.93. However, F_1 scores reveal a big performance drop with the apoptotic class (Figure 3.5), driven by the low 0.5 precision score when predicting the apoptotic label. Precision scores otherwise remain above 0.92 for the other labels.

Performance of the classifier drops when testing against the CRC TME colonic organoid cultures from Qin *et al.* In this case, subsetting just the organoid cells from the organoid cultures (Figure 3.5B), we observe a global accuracy of 0.91. Looking at the classification details (Figure 3.5C) we see a very similar pattern to the SI LGR5 results; with the apoptotic class presenting the lowest F_1-scores (0.6) characterised by a low precision (0.43). Furthermore, the remaining F_1-scores are also lower overall, with only the S-phase and M-phase classes reaching above 0.9.

When no epithelial filter is applied to the dataset and the model performance is tested against all cell types (i.e., including also fibroblasts and macrophages) global accuracy drops down to 0.87. The relatively high global accuracy does not reflect the failure of the classifier to, yet again, identify the apoptotic cells (Figure

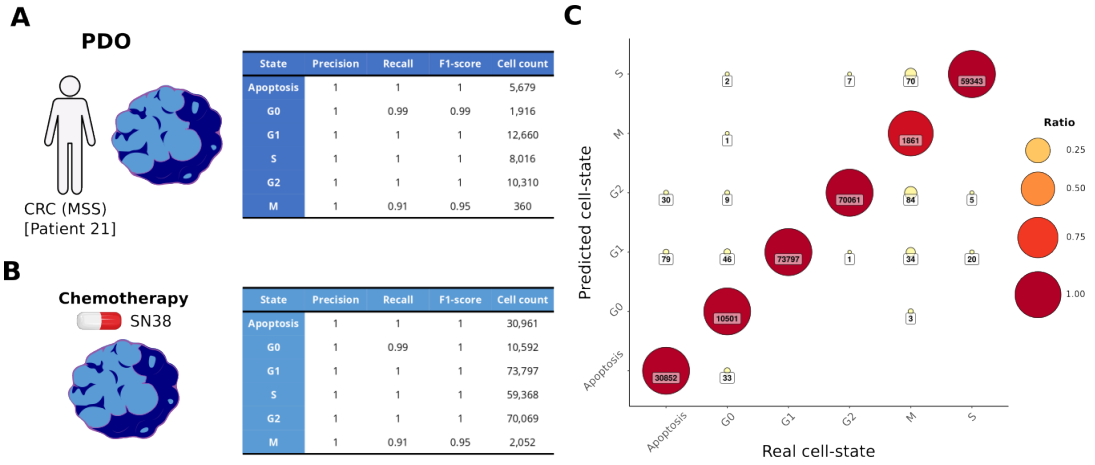


State	Precision	Recall	F1-score	Cell count
Apoptosis	0.5	1	0.67	42,877
G0	0.92	0.92	0.92	365,081
G1	1	0.91	0.95	442,600
S	1	1	1	134,045
G2	1	0.93	0.96	11,590
M	1	0.97	0.98	5,814

State	Precision	Recall	F1-score	Cell count
Apoptosis	0.43	1	0.6	10,213
G0	0.89	0.85	0.87	73,449
G1	0.68	0.91	0.77	44,439
S	1	1	1	184,885
G2	1	0.77	0.87	108,502
M	0.97	1	0.98	9,925

State	Precision	Recall	F1-score	Cell count
Apoptosis	0.32	1	0.48	17,024
G0	0.84	0.80	0.82	165,036
G1	0.71	0.94	0.81	189,249
S	1	1	1	192,255
G2	1	0.67	0.80	264,145
M	0.97	0.99	0.98	10,882

Figure 3.5: 5-marker RF Cell-State Classifier Benchmarks. A-D) Classification reports obtained from running the 5-marker RF classifier against data manually labelled for cell-state from Qin *et al.* [4]. Performance against an intestinal organoid dataset is similar to the training data for the model. Performance against epithelial cells only C) or D) all cell-types from unfiltered cells of colonic heterotypic co-cultures. **E)** Classification matrix from the results in D). Size and colour show predicted to real label ratios, numbers show cell count in each matrix entry.



State	Precision	Recall	F1-score	Cell count
Apoptosis	1	1	1	5,679
G0	1	0.99	0.99	1,916
G1	1	1	1	12,660
S	1	1	1	8,016
G2	1	1	1	10,310
M	1	0.91	0.95	360

State	Precision	Recall	F1-score	Cell count
Apoptosis	1	1	1	30,961
G0	1	0.99	1	10,592
G1	1	1	1	73,797
S	1	1	1	59,368
G2	1	1	1	70,069
M	1	0.91	0.95	2,052

Figure 3.6: 10-marker RF Cell-State Classifier Benchmarks. A) Building a RF classifier with an increased number of markers using data from PDOs achieves better results than the original 5-marker model. **B)** Performance against chemotherapy-treated and untreated PDOs. **C)** Classification matrix from the results in B). Size and colour show predicted to real label ratios, numbers show cell count in each matrix entry. MSS, micro-satellite stable.

3.5D). In Figure 3.5E the classification matrix is used to build a dot plot in which the true labels ("Real state" from gating) are compared against the predicted labels ("Predicted state"), highlighting how a majority of the cells labelled as apoptotic are actually G0 cells, explaining the precision of 0.32 for the former class. There is also some confusion around the G2 cells, as a significant number of these cells are classified as either G0 or G1.

3.3.2 10-marker Model Improves Apoptotic Classification

Given the 5-marker model limitations when resolving the apoptotic class, I implemented a second model using additional PTM antibodies and cell-state markers targeting apoptotic cells. This 10-marker model uses a dataset from Ramos Zapatero & Tong *et al.* [3] wherein heterotypic patient-derived organoid (PDO) cultures from different donors were treated with a spectrum of chemotherapies. This data was generously provided by Dr. Maria Ramos Zapatero.

Results from the updated 10-marker implementation using PDO data show improved performance when compared to the 5-marker model. Using a technical replicates of the training data as test we observe how the apoptotic class gets accurately resolved (Figure 3.6A). When benchmarking the model performance against a

dataset wherein the organoid cells had been treated with SN-38, the active metabolite of the type I topoisomerase inhibitor Irinotecan [188], we observe a global accuracy greater than 0.99. The lowest F_1-scores, at 0.95, were found for the M-phase label (Figure 3.6B). This lowered, yet still accurate, prediction performance is driven by the lower total count of M-phase cells (one order of magnitude smaller than the other classes), hampering the training for that class and resulting in small number of non-apoptotic cells to be miss-labeled as M-phase. In contrast with the 5-marker model results, there is an apparent lack of issues when classifying the apoptotic class, with only 0.35% of true apoptotic cells being mislabelled (Figure 3.6).

3.4 Conclusions

In this chapter I have shown how CyGNAL is an accessible workflow to non-computational users that facilitates data processing and analysis of MC experiments. The computation of EMD and DREMI scores enables a detailed mechanistic description of changes across conditions, wherein changes in the user defined reference allows for differential interrogation of the experimental system.

While the scores themselves can be used to build curated mechanistic models as in Qin *et al.* [4], CyGNAL also incorporates interactive visualisation modules that can automatically plot results. The interactive nature of the visualisation steps, coupled with additional data correlation metrics given during the PCA computation, allows for both exploratory data analysis and (close to) publication grade results generation within a single tool. This same PCA computation presents a straightforward way to summarise changes at the condition level from otherwise information-dense EMD or DREMI heatmaps.

The incorporation of miscellaneous data handling helper scripts in the utilities folder exemplifies how user-provided feedback is paramount, while it also signifies how CyGNAL continuously grows and changes with time. Tools are meant to be used, and that publications by colleagues such as Michelozzi *et al.* [186] employed CyGNAL is a testament to its accessibility.

Originally meant as a simple exercise in curiosity-driven exploration after noticing the correlation between so called PTM and ”cell-state” markers, and empowered by the tediousness of manually gating the datasets in our lab, the RF cell-state classifier has become a convenient tool to automate cell-state labelling of MC datasets in relation to cell-cycle phases.

Albeit a very simple model, the nature of the manual gating process (essentially thresholding on a biaxial space of marker expression) translates well to decision trees, and this is shown in the relatively strong overall model performance. The current implementation however, might struggle to generalise to external datasets, for gating strategies are somewhat of a lab- and individual-specific process.

Where we do observe weak points in the classifier is for those cell-state labels

whose antibody coverage is not great in the model. For example, in the 5-marker RF model, apoptotic cell class precision reaches only 0.32 in the most stringent setting tested (Figure 3.5D). This can be relatively straightforward to address by increasing the number of antibodies targeting that particular state (Figure 3.6B-C), but this strategy can not always be employed as the additional marker would both need to be in the reference data used to train the model and in the query dataset to be labelled. When possible however, as demonstrated by the the 10-marker RF model built, high precision and recall scores are accomplished for all cell-state classes even in the context of cell-cycle disrupting chemotherapy (Figure 3.6B-C).

As described in Chapter 2, both tools are publicly accessible in their respective GitHub repositories.

Chapter 4

Stromal and Oncogenic Regulation of Colonic Stem Cell Polarisation

4.1 Introduction

As presented in Chapter 1, the colonic epithelium is a highly heterogeneous system with multiple specialised cell types. Supported by the *LGR5*$^{+}$ colonic stem cells (CSCs) of the crypt, its homeostatic regulation relies on intrinsic and extrinsic signalling cues, the latter of which predominately come from the stromal compartment. In the context of colorectal cancer (CRC), and under the classical progression model [12], oncogenic mutations targeting *Apc*, *Kras*, *Braf*, *Smad4*, and/or *Trp53* constitute intrinsic cues that are sufficient to induce a highly proliferative crypt-progenitor phenotype, the hyper-proliferative CSC (proCSC) [25]. Therefore, in both the healthy colon and CRC a compartment of epithelial cells is maintained in a stem-like state, although by different mechanisms.

This shared crypt-progenitor phenotype actually represents a broader compartment encompassing more than the canonical LGR5^{+} CSCs; with recent studies describing the existence of the *CLU*$^{+}$ *ANXA1*$^{+}$revival stem cell state (revCSC). Reminiscent of foetal-like states, revCSC has been described as a small and non-proliferative compartment involved in tissue regeneration after injury [30, 29] and suggested as a drug-tolerant persister state in CRC [32].

However, the mechanisms of regulation between the different stem cell states largely remain unclear. Involvement of cell extrinsic cues in the form of stroma-secreted ligands, coupled with the association of the TME with CRC progression, suggest that they must also pay a role in regulating the colonic epithelia. The cell extrinsic cues involve signalling pathways that overlap with those affected by the oncogenic mutations, indicating a competition between intrinsic and extrinsic cues to regulate epithelial polarisation might take place during oncogenesis.

Thus, single-cell omic technologies are perfectly placed to understand polarisation of the epithelial compartment at a broader level and reveal the regulation of cell fates by competing cues.

In this chapter I will explore via scRNA-seq data analysis how cell extrinsic and intrinsic cues co-regulate colonic epithelial fate using a heterocellular organoid culture system with both environmental and oncogenic perturbations. I will first

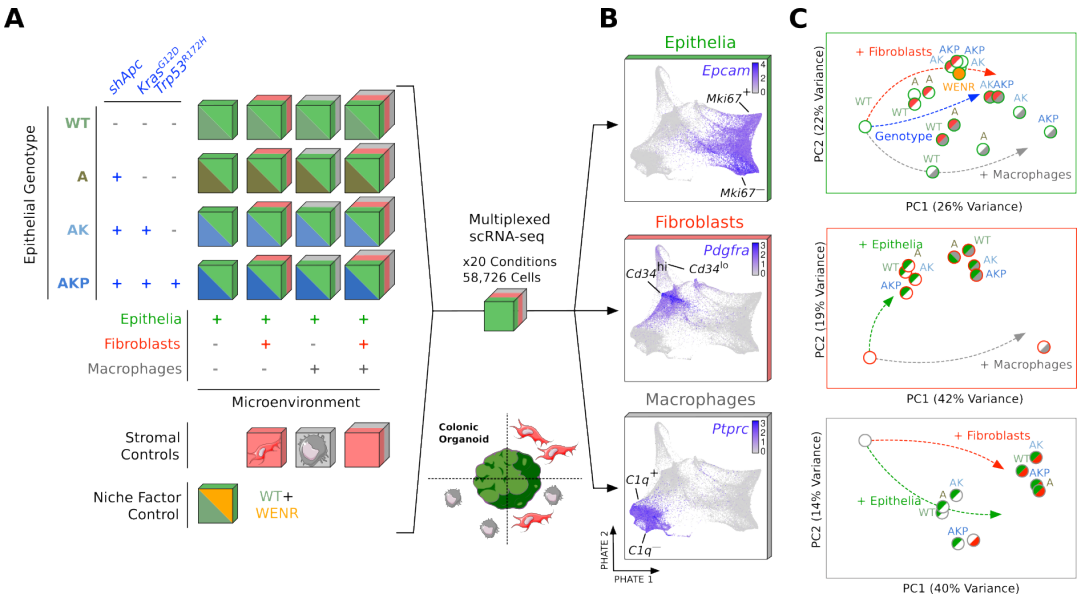


Figure 4.1: Experimental Overview. **A)** Multivariate scRNA-seq experimental design. Recombinant WENR ligands were only present in the niche factor control. **B)** Single-cell PHATE embedding illustrating epithelial cells, fibroblasts, and macrophages. **C)** EMD-based PCA of epithelial, fibroblast, and macrophage transcriptomes. WENR, WNT3A, EGF, Noggin, and R-Spondin-1.

characterise the different populations found in the heterocellular cultures, identify the different epithelial cell states and their compositional changes in response to intrinsic and extrinsic cues. Cellular dynamics approaches will reveal our understanding of the balance regulating the proCSC and revCSC states, and cell-cell communication analysis will suggest putative mechanisms of intercellular regulation. Finally, the findings will be contextualised with the broader literature leveraging published gene signatures.

The work presented here is part of Qin & Cardoso Rodriguez *et al.* [1] (Appendix D), where it is shown accompanied with mass cytometry analyses carried out by Dr. Xiao Qin and whose results validate the scRNA-seq findings and shed light on the mechanisms of epithelia stem cell polarisation in this shared landscape.

To directly compare how CRC oncogenic mutations and stromal cells regulate colonic epithelial differentiation, I performed a multivariate scRNA-seq analysis of wild-type (WT), *shApc* (A), *shApc* and $Kras^{G12D/+}$ (AK), and *shApc*, $Kras^{G12D/+}$ and $Trp53^{R172H/-}$ (AKP) colonic organoid mono- or co-cultures; with colonic fibroblasts and/or macrophages (Figure 4.1A). Fibroblasts are established regulators of intestinal

epithelia [189] and macrophages are the most profuse leukocytes in the colon [190]. A condition with WT organoids cultured with exogenous WNT3A, EGF, Noggin, and R-Spondin-1 (WENR) (commonly used to grow colonic organoids) was included as a defined mesenchymal niche factor control.

Following data acquisition and initial pre-processing steps (see Chapter 2), epithelial cells, fibroblasts, and macrophages were jointly embedded in an integrated space and visualised by PHATE (Potential of Heat-diffusion for Affinity-based Trajectory Embedding) [92]. This embedding resolves the three distinct cell types as shown by expression of levels of canonical cell-type markers (Figure 4.1B). Cell-type-specific transcriptional changes were compared against relevant control monoculture conditions (WT organoids for the epithelial cells) using the EMD score (see 2), and then summarised using PCA (Figure 4.1C). Epithelial transcriptomes are differentially regulated by both CRC mutations (PC1, 26%) and microenvironmental cues (PC2, 22%), with A, AK, and AKP mutations progressively dysregulating their transcriptomic profiles. However, we found fibroblasts can only regulate WT and A epithelial cells (Figure 4.1C). Although WENR ligands are thought to mimic a healthy stromal niche [191], WT organoids + WENR ligands transcriptionally align with AK mutant organoids (not WT+fibroblasts as might be expected), indicating this widely used colonic organoid culture media induces a partial CRC-like transcriptome in WT epithelia (Figure 4.1C). Colonic fibroblast cells resolved into $CD34^{hi}$ and $CD34^{lo}$ subpopulations mimicking *in vivo* stromal heterogeneity [192] (Figure B.1). $CD34^{hi}$ and $CD34^{lo}$ fibroblasts did not differentially regulate colonic epithelia (Figure B.2) and were subsequently treated as a heterogenous mesenchymal population. Bone marrow macrophages on the other hand presented as a continuum of cells aligned along and axis of inversely correlated expression of complement genes (like *C1q*) and *Hmox1*, see 4.1B), possibly indicating inflammation-related roles to be a major driver of heterogeneity within the macrophage cells [193]. However, it was found that fibroblast and macrophage transcriptomes and compositional make-up were only regulated by co-culture with heterotypic cells but not altered by epithelial genotypes (Figures B.1, B.3).

4.2 Organoids Recapitulate Colonic Epithelial States

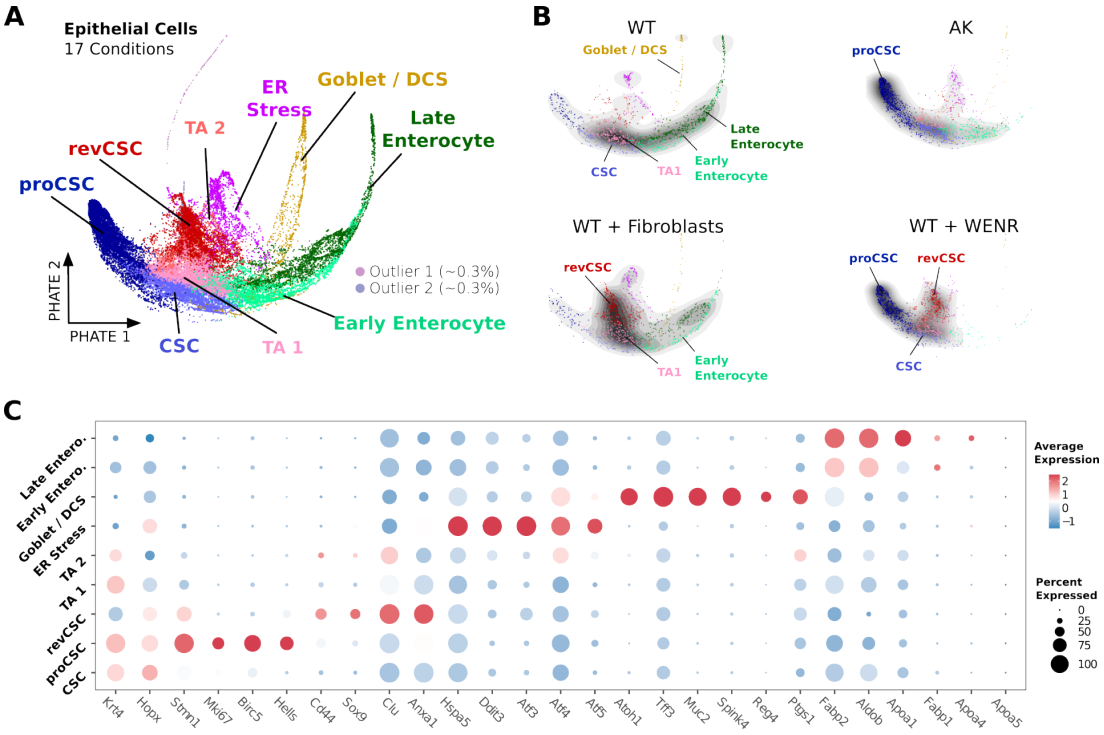


Figure 4.2: Recapitulation of Colonic Epithelial States. **A)** PHATE embedding of epithelial cells from all organoid conditions, coloured by cell-type clusters. **B)** Single-cell PHATE embeddings of epithelial cells from WT, WT+Fibroblasts, WT+WENR, and AK organoids coloured by cluster and overlaid with single-cell density. **C)** Expression of *bon-fide* epithelial markers in agreement with cluster designations. Colour is scaled average gene expression by cluster, size is ratio of cells in cluster with detected marker expression. CSC, colonic stem cell. proCSC, hyper-proliferative CSC. revCSC, revival CSC. DCS, deep crypt secretory (cell). TA, transit amplifying (cell).

Epithelial cells from all conditions were integrated by reciprocal PCA (RPCA) [147], projected onto a shared PHATE embedding, and clustered into multiple cell-fates, including stem populations, transit amplifying (TA) cells, cells under ER stress, goblet and deep crypt secretory (DCS) cells, and early or late enterocytes (Figure 4.2A).

While this integrated space presents a continuum of cells, density plots of 4 extremes in our experimental design matrix point towards some degree of polarisation (Figure 4.2B). The WT monoculture control spans a broad range in the embedding space and shows high density in the CSC to Enterocyte and Goblet/DCS differentia-

tion axes. The WT cocultured with fibroblasts appears to show the highest density of cells around the revCSC state, whereas the AK monoculture is densest around the proCSC. Finally, the condition with exogenous WENR ligands seems to polarise both towards proCSC and revCSC.

The multiple epithelial compartments where identified and associated with the relevant clusters based on their expression of canonical markers of selected colon cell epithelia cell-types (see Sup. Table C.1 for more epithelial marker genes). Expression of these genes on the WT monoculture control reveals how the system recapitulates the basal (stem and TA), secretory, and absorptive compartments (Figure 4.2C). The stem compartment appears distributed along several clusters (proCSC, CSC, revCSC) and extends towards the TA cells. A state characterised by a clear ER stress response gene expression signature lays adjacent to the stem and TA compartments (Figures 4.2A and C).

4.3 Mutations and Fibroblasts Polarise Epithelia towards Distinct Cell Fates

Compositional analysis via differential abundance / differentially abundant (DA) testing [103] is used to identify and quantify effects of perturbations on a system (see Chapter 2 for more details). DA was thus used to determine the changes induced by stromal and oncogenic cues compared to the WT mono-culture organoid baseline, revealing that fibroblasts and CRC mutations have markedly different effects on epithelial cell-fate determination (Figure 4.3).

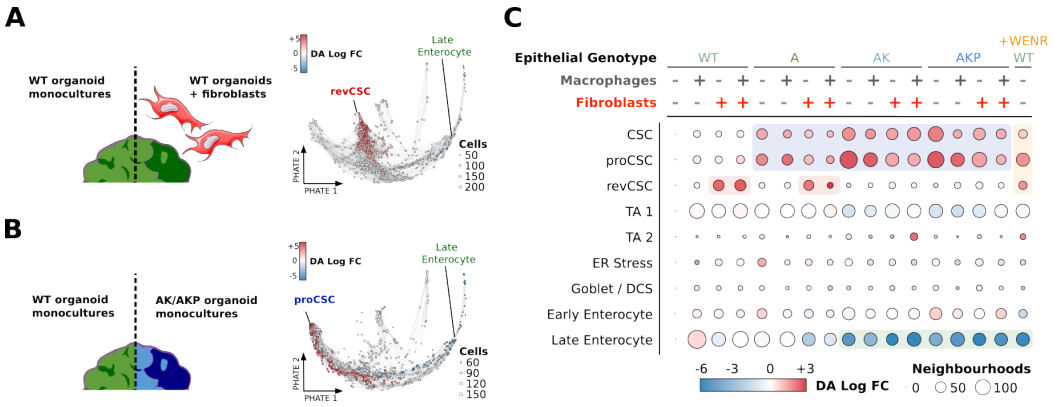


Figure 4.3: DA Reveals Oncogenic and Stromal CSC Polarisation. A) Epithelial DA neighbourhoods in WT organoid and fibroblast co-cultures compared to WT organoid mono-cultures. Colour indicates log fold-change, size indicates number of cells in the neighbourhood. **B)** Epithelial DA neighbourhoods in AK and AKP organoid mono-cultures compared to WT organoid mono-cultures. Colour indicates log fold-change, size indicates number of cells in the neighbourhood. **C)** Overview of per-cluster epithelial DA changes across organoid cultures. Colour indicates log fold-change, size indicates number of neighbourhoods. DA, differential abundance / differentially abundant. FC, fold-change.

Fibroblasts enrich for the revCSC population characterised by high expression of epithelial progenitor genes *Clu*, *Sox9*, *Cd44*, and *Cldn4* (Figures 4.3A, 4.4). In contrast, A, AK, and AKP mutations progressively polarise epithelia towards a hyper-proliferative proCSC state (Figure 4.3B). proCSCs express *EphB2*, *Birc5* (*Survivin*), *Lrig1*, *Hmgb2*, *Anxa1*, and *Rrm2*. proCSC are also highly mitotic, expressing *Stmn1*$^{+}$, *Mki67*$^{+}$, and *Ccnb1*$^{+}$ (Figures 4.2C, 4.4).

Both revCSC and proCSC are present in WT organoids at low levels alongside traditional *LGR5*$^{+}$ CSCs, and these cells were found to also be enriched by A, AK,

Figure 4.4: Curated Differential GEx Analysis of Epithelial Cells. Heatmap of gene signatures curated from the literature and DE analysis. Columns are aggregated by clusters and colour-annotated with metadata labels. Gene colours represent scaled gene expression. GEx, gene expression.

and AKP genotypes, but to a lesser extent than proCSC (4.3C).

While the DA method employed essentially works at a pairwise level, I aggregated results from multiple comparisons across the experimental matrix (Figure 4.3C) to show how how fibroblasts can only induce revCSC in WT and *shApc* epithelia, but not when cells contain both *shApc* and $Kras^{G12D/+}$. Conversely, proCSCs are enriched in all A, AK, and AKP organoids irrespective of fibroblasts or macrophages; suggesting oncogenic mutations are dominant over microenvironmental signalling. WENR ligands polarise WT epithelia towards all stem and TA cell-types, with very few cells retaining secretory or absorptive identities (Figures 4.3C, 4.2B). While macrophages can alter epithelial gene expression (Figure 4.4), macrophages do not regulate the abundance of epithelial cell-types (Figure 4.3C).

In summary, multivariate scRNA-seq revealed that fibroblasts, CRC mutations, and WENR ligands polarise epithelia towards a de-differentiated progenitor state – with fibroblasts and oncogenes inducing distinct revCSC and proCSC fates.

4.4 Epithelial Dynamics Suggest Transitional Regulation of revCSC

To understand the nature of the epithelial polarisation observed I leveraged methods that infer transcriptional dynamics from the static snapshots found in the experimental setup.

The CCAT metric is a measure of cellular pluripotency [110] completely independent of cluster and other metadata designations (see Chapter 2). Paired with RNA velocity information [100], both metrics revealed how the stem clusters present the highest pluripotency scores and act as origin for the RNA velocity stream embeddings (Figure 4.5A).

Contrary to proCSC, revCSC shows the lowest pluripotency score of all stem and TA clusters, and overall CCAT is able to position the clusters along the expected stem to differentiated states trajectory (Figure 4.5B). RNA velocity [100] vector lengths were used a metric for the rate of transcriptional change (see Chapter 2) and reveal how, while the WENR organoids show significantly decreased rates of change around the proCSC compartment, AK organoids present a 2-fold reduction of velocity vector lengths across all epithelial compartments (Figure 4.5C-D).

The RNA velocity information was then used to infer transitional processes and trajectories with CellRank [101]. Determination of initial and terminal macro-states in the 4 conditions from Figure 4.2B consistently identifies proCSC as the source of transitional processes within the system (Figure 4.5E). In the WT mono-culture control the expected differentiation trajectories are recovered, whereas polarisation towards revCSC by fibroblasts and WENR appears to be a transitionally driven event from the nearby stem and TA states. In AK organoids the limited amount of transitions detected are towards the remnants of the secretory and absorptive populations, yet the proCSC still appear only as source and not a sink (Figure 4.5E), altogether suggesting that oncogenic mutations reduce epithelial plasticity.

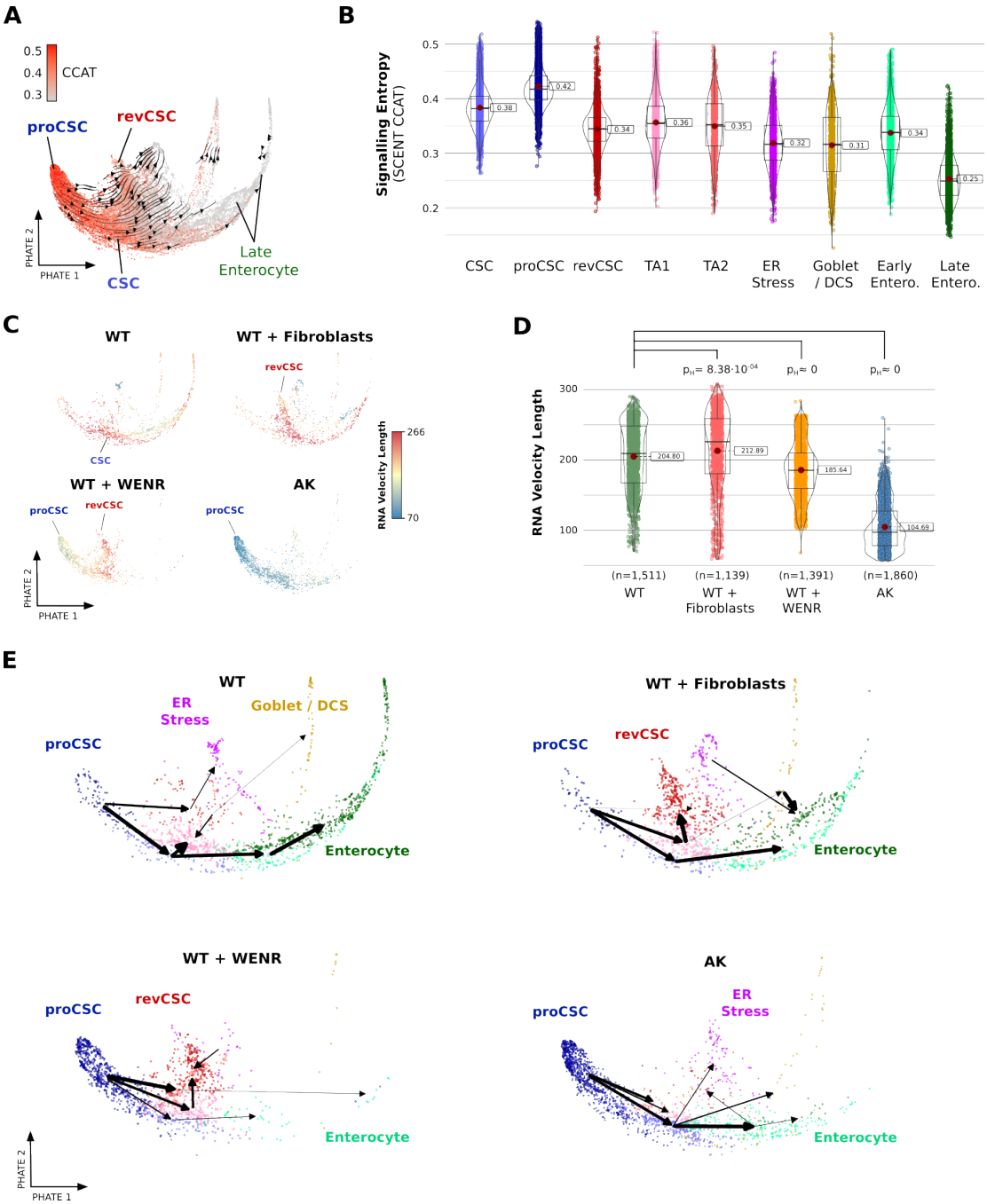


Figure 4.5: Cellular Dynamics of Epithelial Polarisation. A) Epithelial PHATE coloured by CCAT score and overlaid with RNA velocity streams (arrows). **B)** Distribution of CCAT scores per epithelial cluster. **C)** Epithelial PHATE coloured by RNA velocity vector lengths. **D)** Distribution of RNA velocity vector lengths per organoid condition (Games-Howell pairwise test with Holm-adjusted p-values). **E)** Directed PAGA plots depicting transitions from initial to terminal macrostates. Colour denotes epithelial cluster, arrow width represents aggregate RNA velocity flows.

4.5 Oncogenic Mutations Disrupt Fibroblast to Epithelia Signalling

As epithelial differentiation cannot be regulated by fibroblasts in the context of *shApc* and $Kras^{G12D/+}$ (Figures 4.3C, 4.4), I hypothesised oncogenic mutations might disrupt stromal-epithelial signalling. To test this, I performed cell-cell communication analysis with CellChat [49] of WT, A, AK, and AKP organoid-fibroblast co-cultures.

By aggregating incoming and outgoing communication probabilities (a measure of the degree of expression for ligands and receptors belonging to predicted cell-cell interactions) I observed high levels of 'outgoing' signalling from fibroblasts (Figure 4.6A). By contrast, WT epithelia display a dominant 'incoming' signalling potential (Figure 4.6A). This dichotomy suggests that heterocellular signalling in the healthy colon is largely unidirectional from fibroblasts to epithelial cells. The revCSC and the transcriptionally similar TA clusters are responsible for much of the 'incoming' signalling potential of WT epithelia, indicating these states are hyper-sensitive to cell-extrinsic regulation by fibroblasts. In contrast, proCSC are the least receptive of all epithelial cells, suggesting proCSC are more reliant on cell-intrinsic signalling (Figure 4.6A).

An overview of stroma-derived interaction changes on the epithelial states across genotypes revealed that fibroblasts communicate with the organoids both by juxtacrine and paracrine interactions (Figure 4.6B). Not only is there a loss of predicted interactions in AK and AKP cells compared to WT organoids (Figure 4.6A), but there are also some signalling pathways that appear missing on the cancer organoids. For example, WT and A organoids show intact NRG1, EREG, IGF, and TGF-β signalling with fibroblasts, but these cell-cell interactions are undetectable in AK and AKP cells. These predicted signalling pathways can be cross-referenced with the components of the WENR-enriched media to suggest some ligands as WT homeostatic regulators, such as WNT5A, SEMA3A, TGF-β1, TGF-β2, IGF, NRG1, EREG, and OSTP (encoded by *SPP1*).

The observed breakdown in fibroblast to epithelia communications might partially be explained due to the downregulation of epithelial signal receptors in AK

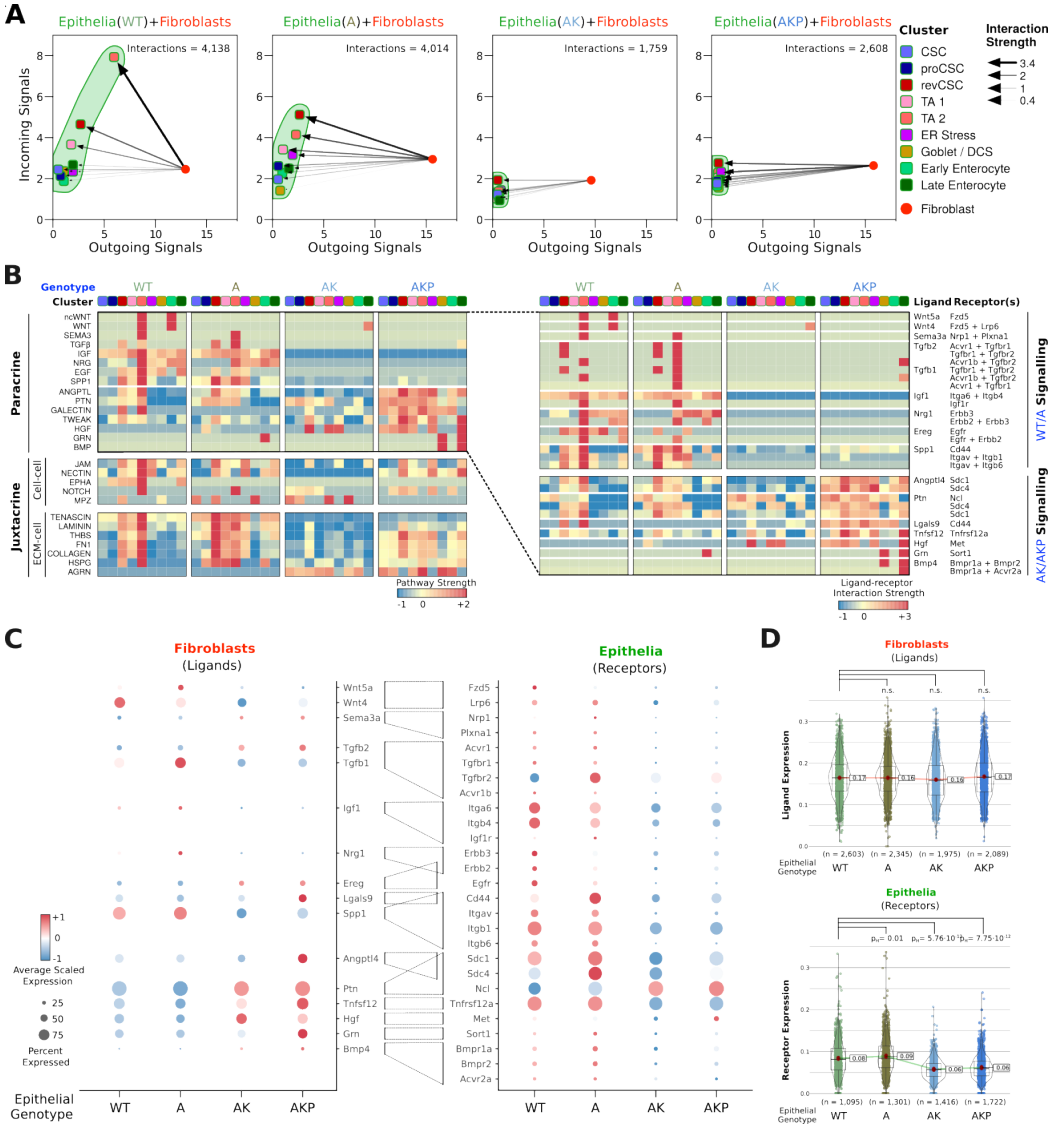


Figure 4.6: Oncogenic Mutations Disrupt Stromal Communication. A) Outgoing and incoming communication probability (interaction strength) from fibroblasts to epithelia across organoid genotypes. Arrow size denotes aggregate fribroblast-to-epithelia communication probability. **B)** Paracrine and juxtacrine communication summarised at the pathway and ligand-receptor interaction level. **C)** Expression of individual ligands (expressed by fibroblasts) and receptors (expressed by epithelia) across organoid genotypes. Colour shows average scaled expression, size is ratio of cells with detected expression. **D)** Aggregate UCell [5] scores for ligand expression on fibroblasts and receptor expression on epithelia across organoid co-cultures (Games-Howell pairwise test with Holm-adjusted p-values, n.s not significant).

and AKP organoids (Figures 4.6C-D), while ligand expression by the fibroblasts remains unchanged (Figure 4.6D).

4.6 Characterisation and Relevance of proCSC and revCSC Identities

As described in Chapter 2, literature signatures for diverse epithelial stem states and transcriptions targets of key signalling pathways were curated (see C.9) and compared against our scRNA-seq data using UCell [5]. This analysis revealed how the fibroblast-induced revCSC are indeed transcriptionally similar to ”foetal” [158, 159] and “revival” stem cells [30] of the intestinal epithelia (4.7A).

The previously described association between revCSC and Yap and TGF-β was also recovered by signature correlation, further validating the identity of the revCSC cluster. This observation, together with the cell-cell communication results, provided with the initial targets to pursue the mechanistic discovery of master regulators of the different stem states [1].

In addition, proCSCs are transcriptionally comparable to stem cells observed in mouse and human CRC (Figure 4.7A), showing a clear link with actively proliferating stem cell populations. CSC gene signatures are less common in CRC (Figure 4.7A) and more closely resemble general pan-stem states.

This link between our mouse organoids and CRC patient data was further explored by comparing the murine organoids with aggregated scRNA-seq data from several CRC cohorts in Joanito *et al.* [6]. This resource contains both tumour and normal tissue samples that could be resolved on an integrated space of their scRNA-seq profiles (Figure 4.7B). After cross species integration and projection of our mouse organoid data, one can non-quantitatively observe that WT organoid cells align with normal tissue, whereas AK organoids align with cancer samples (Figure 4.7C). See Chapter 2 for further details on this process.

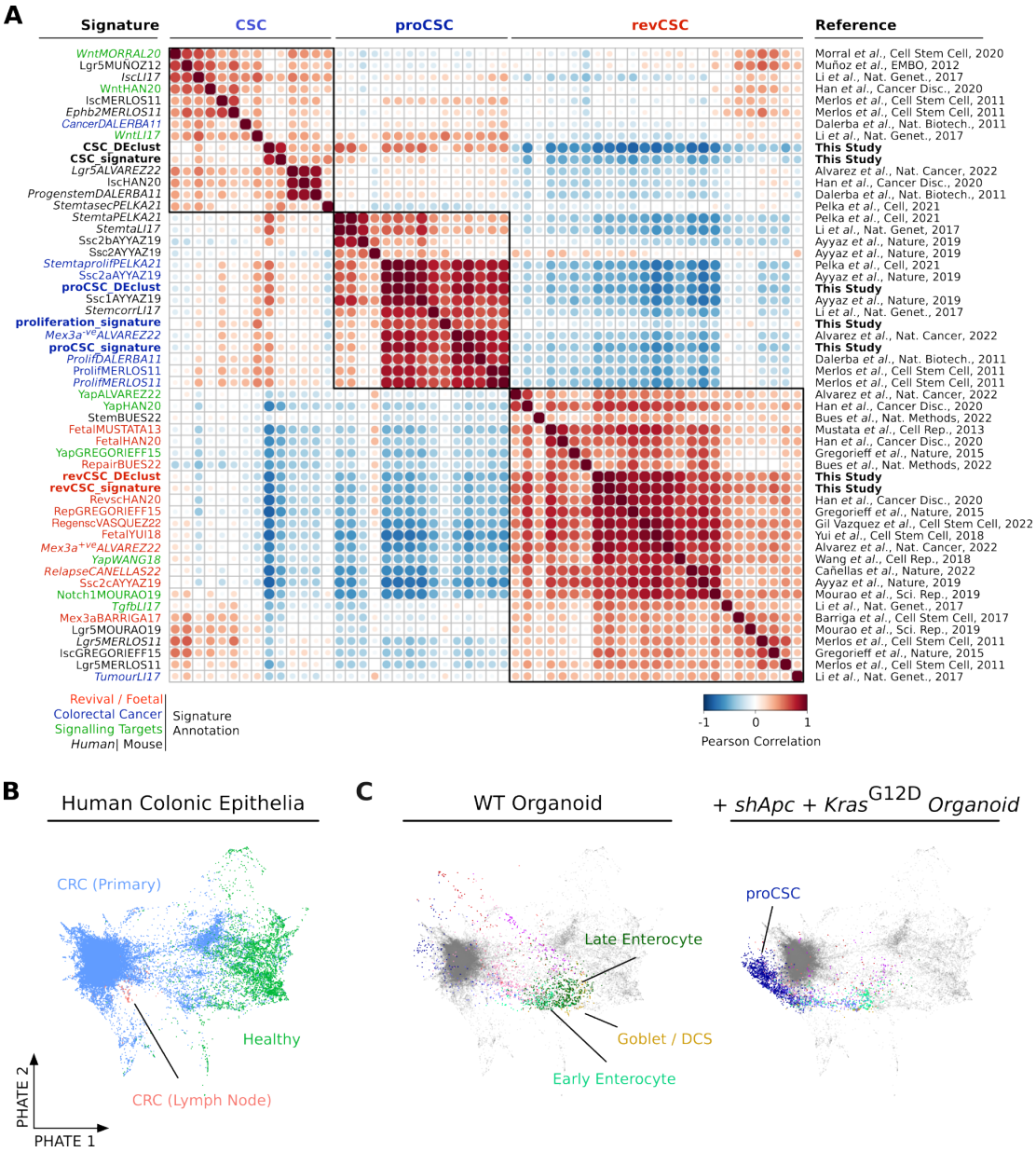


Figure 4.7: Epithelial Stem Cell Signature Comparison. A) Comparison of CSC, proCSC, and revCSC gene signatures identified in this study with published stem cell and signalling signatures. Colour denotes Pearson correlation of UCell [5] scores. **B)** Epithelial PHATE of integrated CRC cohort from Joanito *et al.* [6]. Colour marks sample type annotation. **C)** Projection of our murine WT and AK organoid data on human PHATE embedding.

4.7 Conclusions

In this chapter I have shown how scRNA-seq can be used to dissect a cell-type-specific understanding of heterocellular organoids in cancer. I have provided with an in-depth description of colonic epithelial differentiation and polarisation of its stem compartment. Furthermore, *in silico* predictions regarding the mechanisms regulating these processes can also be formulated, which can be (and have been in Qin & Cardoso Rodriguez *et al.* [1]) functionally validated using alternative single-cell *omic* approaches (MC).

On the unperturbed control, WT organoid monocultures recapitulate canonical differentiation from stem and basal states towards secretory and absorptive compartments. However I was unable to discern between discrete states within the secretory populations, with data from similar murine small intestinal organoids revealing the same observation, suggesting a putative limitation of the organoid model when compared to the *in vivo* setting.

The finding of a heterogeneous stem compartment that can be so drastically polarised by stromal and oncogenic perturbations immediately stands out as the central observation of this study, revealing that fibroblasts and oncogenic mutations induce distinct epithelial stem cell-fates in colonic epithelia. I found that fibroblasts, potentially through the secretion of signalling ligands linked with WNT and TGF-β1, polarise epithelia towards slow-cycling *Clu*$^{+}$ revCSCs. In contrast, simultaneous APC-loss and oncogenic KRAS^{G12D} collaboratively block cell-extrinsic regulation of epithelial plasticity by interrupting stromal-epithelial communication, and polarise the organoids towards the hyper-proliferative proCSC state. By comparing the transcriptomic profiles of the stem states with the literature, I was able to validate both their identity and the link between revCSC and TGF-β1 and YAP signalling, while also validate the organoid model as a whole.

The addition of WENR-enriched media revealed that exogenous WNT and EGF ligands can polarise the epithelium towards all stem states, at the expense of the differentiated cell states. In Qin & Cardoso Rodriguez *et al.* [1] we experimentally demonstrate that CRC organoids can still access revival stem cells, but this requires

high cell-extrinsic activation of YAP via TGF-β1 in parallel with reduced PI3K signalling.

CCAT scores have been shown to be a powerful metric to determine putative cellular pluripotency scores, which can then be used in establishing cellular identity and inform dynamic transitional processes. The CCAT pluripotency metrics and RNA velocity results are orthogonal methods that both paint a shared picture of competing transition and differentiation. In this shared landscapes proCSC gives rise to differentiated states in the unperturbed organoids, but can be polarised towards alternatives fates like revCSC (via extrinsic cues) or trapped in the proCSC state by oncogenic mutations.

These results demonstrate that colonic epithelia exist on a continuous differentiation landscape where oncogenic mutations and stromal cues compete for epithelial identity. However it appears that oncogenic mutations eventually dominate extrinsic cues by blocking the stromal regulation of cell-fate plasticity.

Chapter 5

Data-driven Landscapes of Colon Epithelial Plasticity

5.1 Introduction

More than 60 years ago, Conrad H. Waddington illustrated the process of an epigenetic landscape where pluripotent cells would roll down into valleys of terminally differentiated states [129]. Albeit a powerful image of developmental biology, his effort and subsequent ones since then have mostly been of a rather subjective and artistic nature. However, reconstructing such landscapes from real biological data is not an untenable task anymore, as *omic* profiles from single-cells can be embedded together and mapped onto a 3D space sculpted by cellular pluripotency metrics [130].

However, none of those methods appear to leverage embeddings able to capture transitional processes and global structure. Furthermore, such a Waddington-like landscape would need to be informed by features at multiple levels: with a coarse feature informing overall elevation, and local information determining the presence of troughs and valleys, thus shaping the repertoire of likely downhill transitions.

Here I propose a novel method to generate such data-driven Waddington-like landscapes using; 1) embeddings that capture global structure (PHATE [92]), 2) a cellular pluripotency metric to derive coarse landscape elevation (CCAT [110]), and 3) RNA velocity metrics to capture local transciptomic changes (scvelo [100]) that inform state accessibility (Figure 5.2A).

This work has been published as part of Qin & Cardoso Rodriguez *et al.* [1], and the code to compute the VR score and generate the landscapes is publicly available as a Jupyter Notebook on `github.com/TAPE-Lab/Qin-CardosoRodriguez-et-al/blob/main/Figure7_S7/Landscape.ipynb`.

5.2 The Valley-Ridge Score

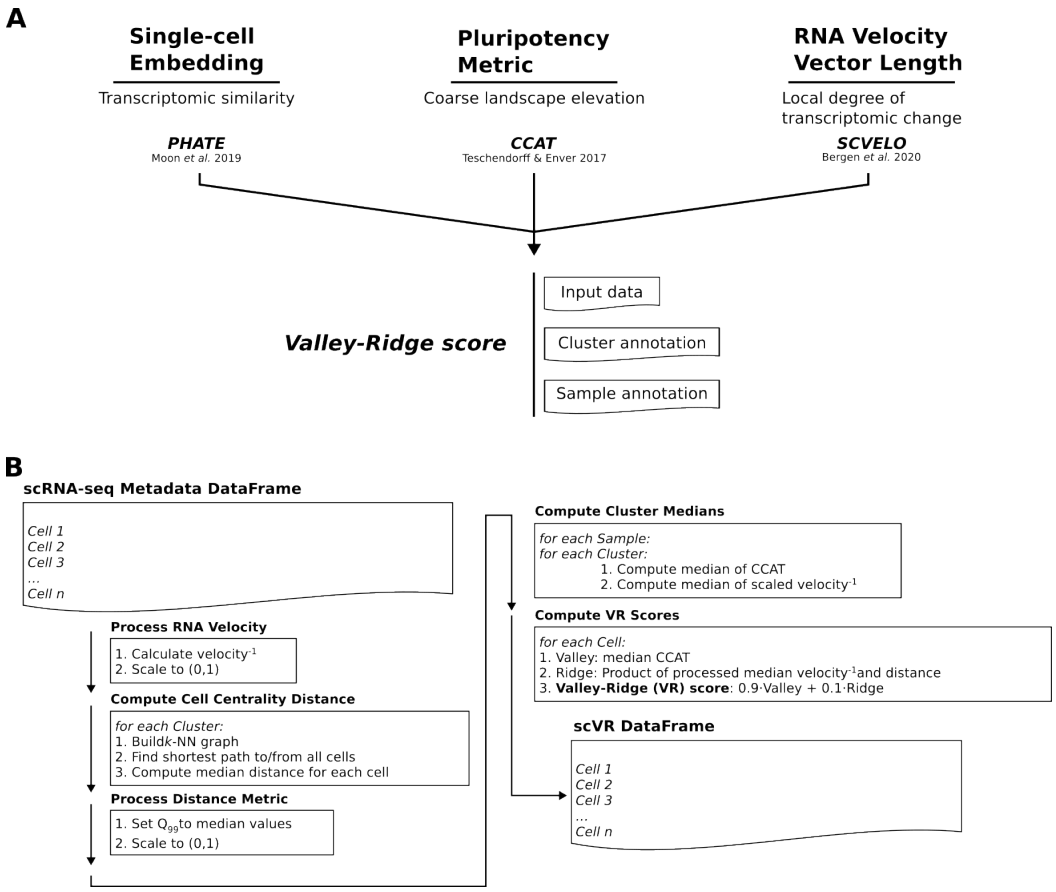


Figure 5.1: Workflow for Calculating VR Scores from scRNA-seq Data. **A)** VR scores leverage a low-dimensional embedding and are computed from pluripotency and RNA velocity metrics. **B)** Computation of VR scores incorporates global and local components as a weighted sum. VR, valley-ridge. Q_{99}, 99th quantile.

Following with the geographical analogy, PHATE space acts as the *longitude* and *latitude* coordinates whereas we need to define a new metric that combines both CCAT scores and RNA velocity vector lengths. This metric has been called the Valley-Ridge (VR) score, in reference of its two components that respectively inform macro-level and hyper-local features of the landscape (Figure 5.2A).

While these two metrics have already been discussed previously in this work, here is a small summary of what they entail. CCAT has been defined as an estimate for a cell's Signalling Entropy Rate, which has been shown to be a robust metric for cellular pluripotency [110, 130, 112]. RNA velocity vector lengths are the modulus of the inferred RNA velocity vectors as determined by a cell's ratio of spliced

and unspliced mRNA, thus measuring the overall rate of transcriptomic change undergone by a cell.

Detailed information on the definition and computation of the VR score can be found in Chapter 2. In brief, the VR score is a cellular metric computed on a per sample and cluster labels and is defined as the weighted sum of the two components: CCAT signalling-entropy [110] and RNA velocity vector length [100] (Figure 5.2B). At a cluster's centre, the VR score is solely determined by the median CCAT. However, the VR scores at the cluster periphery are augmented by weighting the inverse of RNA velocity component and the scaled distance from the cluster centre to model rates of local transcriptional change. We use the inverse of the velocity vector length so that transitions substantiated by high RNA velocities do not locally increase landscape elevation at a cluster's boundary, with the opposite happening for low velocity cells.

This method thus reconstructs a data-driven estimate of Waddington-like landscapes where the overall altitude captures the differentiation potential of a cell population, with the valley-ridge topology delineating local plasticity and cell-state availability.

5.3 Landscapes of Colonic Epithelia Cell-Fate Plasticity

Having been described in Chapter 4 and in Qin & Cardoso Rodriguez *et al.*, the heterocellular murine colonic organoid system represents a suitable candidate to test the VR landscapes. This system consists of colon epithelia organoids increasingly accumulating canonical CRC oncogenic mutations, and with various combinations of microenvironmental perturbations including a stromal component (Figure 5.2A).

The cellular dynamics of this system suggest stromal cues polarise colonic epithelia towards a slow-cycling revCSC state, comprising lower pluripotency potential than other stem states. There is also some loss of terminally differentiated states by stromal cues, especially around the absorptive compartment, but this de-differentiation was not as pronounced as in CRC organoids. Oncogenic mutations polarise epithelia to the proliferative and highly pluripotent (as determined by CCAT) proCSC state. Furthermore, RNA velocity vector lengths in CRC organoids were greatly reduced when compared to the other genotypes (Figure 4.5C-D), suggesting that normal transitional processes within the epithelia are impeded by oncogenic mutations. The VR score is a way of visualising all of these processes at once by generating a purely data-driven VR landscape reminiscent of Waddington own's drawing.

When WT colonic epithelia are projected onto this embedding, stem cells occupy high positions in the landscape, with TA cells descending into a central valley before diverging into terminally differentiated secretory and absorptive cells (Figure 5.2B). When WT epithelia communicate with fibroblasts, the TA valley erodes as cells access revCSC (Figure 5.2B). In contrast, CRC mutations *shApc* and $Kras^{G12D/+}$ re-sculpt the entire landscape, trapping most cells in the proCSC fate by restricting their differentiation potential (Figure 5.2B).

This landscape projection exemplifies the VR score profile of cellular states such as proCSC, which are highly pluripotent (Figure 4.5B), yet static in terms of rate of transcriptional change (Figure 4.5C). proCSC sates appear as high elevation tarn-like features, surrounded by an obstructive ridge that symbolises the low likelihood of

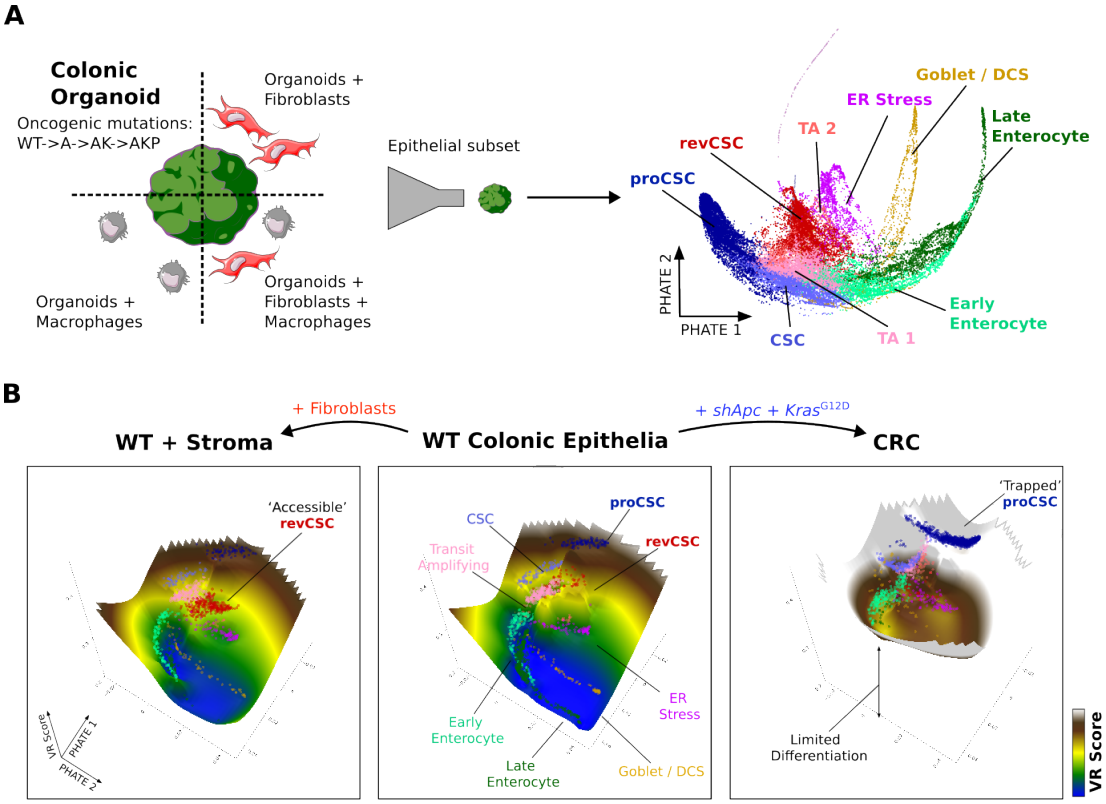


Figure 5.2: Fibroblast- and Oncogene-driven Waddington-like Single-cell Landscapes. **A)** Epithelial cells from the heterocellular CRC organoid model system are used to compute VR scores. **B)** Integrating PHATE and Valley-Ridge (VR) score enables Waddington-like landscapes of scRNA-seq data, illustrating processes of CSC polarisation. Landscape colour denotes VR elevation, dot colours represent epithelial clusters.

transition towards surrounding states. VR landscapes therefore enable us to visualise how proCSC are a stem cell (high in Waddington space) that rarely differentiate (trapped in a tarn).

See Chapter 2 for details on the methods used to interpolate the VR scores into a surface and the pipeline to generate the VR landscapes (Figure 2.3).

5.4 Conclusions

The VR score presented here synthesises two orthogonal metrics (signalling entropy rate and transcriptomic rate of change) that when combined are very useful in visualising transitional processes and plasticity of a system. The multi-scale nature of the its components, with CCAT determining coarser cluster-level features and RNA velocity vector lengths more local inter-cluster transitions, proves useful when reconstructing data-driven Waddington-like landscapes.

When applied to murine organoid perturbation system described in Chapter 4, the VR landscapes depict a picture of a shared differentiation that can be traversed through cell-extrinsic ligands or cell-intrinsic oncogenic mutations. In particular, the increased availability of revCSC in the presence of stromal ligands (Figure 4.3) can also be observed on the VR landscapes (Figure 5.2B). Furthermore, the collapse of stromal-to-epithelial communication in cancer organoids (Figure 4.6A) and their lack of revCSC polarisation (Figure 4.3C) is reflected in the tarn-like topology of the AK VR landscapes, where the bulk of the organoid appears trapped in the proCSC state.

By combining the VR score computation and landscape projection into a single easy to use notebook, I have laid the foundation towards future packaging and deployment of this tool as an interactive service. By the name of VRland, this tool is currently available as an annotated Jupyter Notebook in the code repository for Qin & Cardoso Rodriguez *et al.* [1] (github.com/TAPE-Lab/Qin-CardosoRodriguez-et-al).

Chapter 6

Knowledge Graphs for Cell Communications

6.1 Introduction

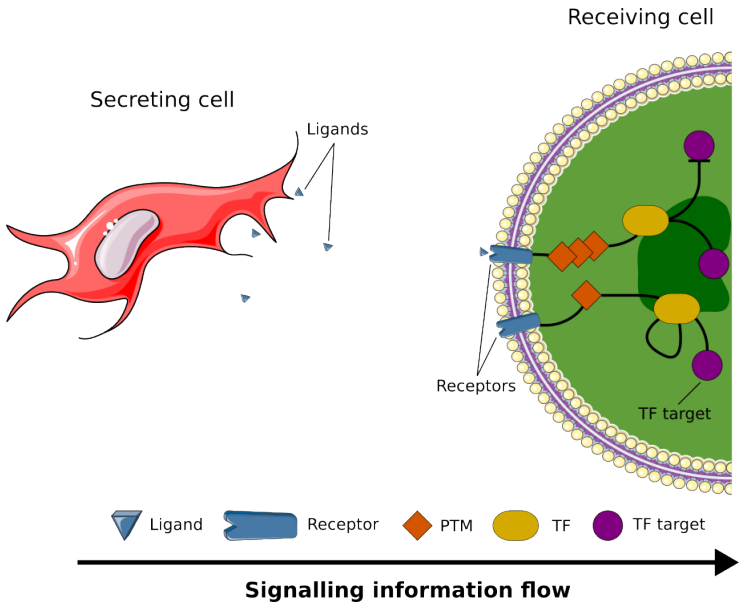


Figure 6.1: The Directed Nature of Inter- and Intra-Cellular Communications. Secreting cells interacting with receiving cells via intercellular ligand-receptor interactions, which can then trigger intracellular PTM cascades and gene-regulatory networks. PTM, post-translational modification. TF, transcription-factor.

Cellular signalling involves a complex series of directed and hierarchical [194] signal transduction cascades between molecules that dictate a cells response to extrinsic and intrinsic cues. In the context of inter-cellular paracrine communication, a secreting cell produces a series of ligands that are captured by receptors on a receiving cell. The receiving cell then might engage in an intra-cellular signal transduction cascade orchestrated by PTMs, such as the MAPK cascade [195]. These cascades regulate gene expression downstream of active transcription factors. With overlapping pathways, feedback loops, and complex settings with multiple cells engaging in symmetrical or non-symmetrical communications, there is nonetheless a directional causality-driven signalling information flow (Figure 6.1). This directional nature can be measured in terms of graph hierarchy scores, and to aid with that purpose I have developed a python package to compute such scores (Appendix A).

The physical interactions between molecules are often represented as a network of genes, proteins or even PTMs, described in the manner of a knowledge graph (KG). These network representations have been extensively explored to model both intra-

and inter-cellular communications, but to date they are not consistently analysed using methods that leverage the underlying directed and hierarchical nature of signalling processes, often either treating the graph as undirected or analyzing pairwise relationships between feature detection metrics (such as gene expression) [196, 121].

The field of directed cellular interaction databases already presents with some established curated resources like OmniPath [7], with a growing number of methods attempting to model communication in a directed manner [132], describing cell-cell interactions [197, 198], and even data-driven *de novo* generation of signal transduction networks [122].

In this chapter I propose a novel approach for assembling gene-gene graphs that capture cellular communication by leveraging KG embedding approaches, which would allow for the encoding of the original directed KG into a simpler non-directed format amenable to downstream analysis and data projection. I aim to project single-cell *omic* profiles into the assembled KGs, thus treating the cells as signals on a gene graph. The resulting signals can then be considered as another single-cell *omic* view of the cells, and used to generate new embeddings or be compared against their gene expression profiles.

This work was conducted in collaboration with Prof. Smita Krishnaswamy and Aarthi Venkat at Yale University, under the Yale-UCL Exchange Programme (`https://www.grad.ucl.ac.uk/yale-ucl/`).

6.2 A Knowledge Graph for Ligands, Receptors and TF Targets

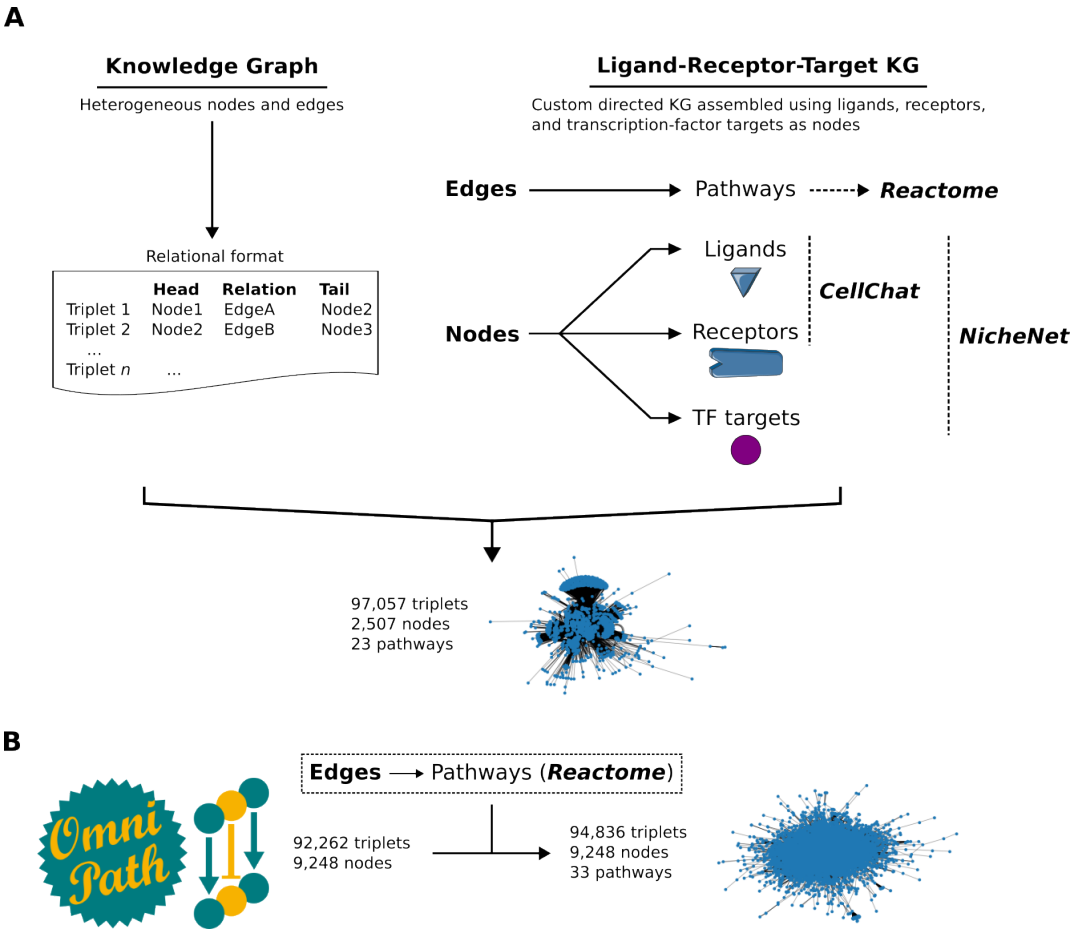


Figure 6.2: Assembly of KGs for Cell Communications. A) Public databases are used to assemble a custom KG of ligands, receptors and TF targets. **B)** Tabular OmniPath [7] repository can also be assembled as a comparable KG. KG, knowledge graph. LRT-KG, ligand-receptor-target KG.

Literature information on cell communication interactions is commonly found in the form of databases used for cell-cell communication analyses, and not in a directed graph format. Therefore I assembled a custom *kg* from public databases and compared it with OmniPath [7], an existing curated repository of directed inter- and intra- cellular signalling interactions. More details on this process can be found in Chapter 2.

I gathered information from the CellChat [49] and NicheNet [178] databases to assemble a directed KG wherein nodes are genes for ligands, receptors or

transcription-factor (TF) targets (Figure 6.2). This KG aims to capture inter- and intra-cellular communication; with ligand and receptor nodes describing the relationship between interacting cells, and the TF targets capturing cellular states and response to stimuli.

Following the ubiquitous triplet format, I thus encoded the graph as a relational database where pathways from Reactome [118] were used to annotate and relate the different gene nodes (Figure 6.2).

The resulting ligand-receptor-target KG (LRT-KG) has over 2,500 nodes linked by interactions belonging to 23 distinct pathways. To validate broad-scale graph characteristics this custom graph was compared against the OmniPath resource. The OmniPath database has multiple layers of relational information between genes (and other molecules such as PTMs), including directionality, supporting evidence, and functional information on the nature of the interaction (i.e. activation or inhibition of receiving interaction member) [7]. Assembled in the same manner as the LRT-KG object, the OmniPath graph presented with a higher number of gene nodes and pathways but comparatively less interactions and a lower hierarchy score than the LRT-KG (Figure 6.2B and Table 2.3).

6.3 KG Embeddings Preserve Graph and Biological Information

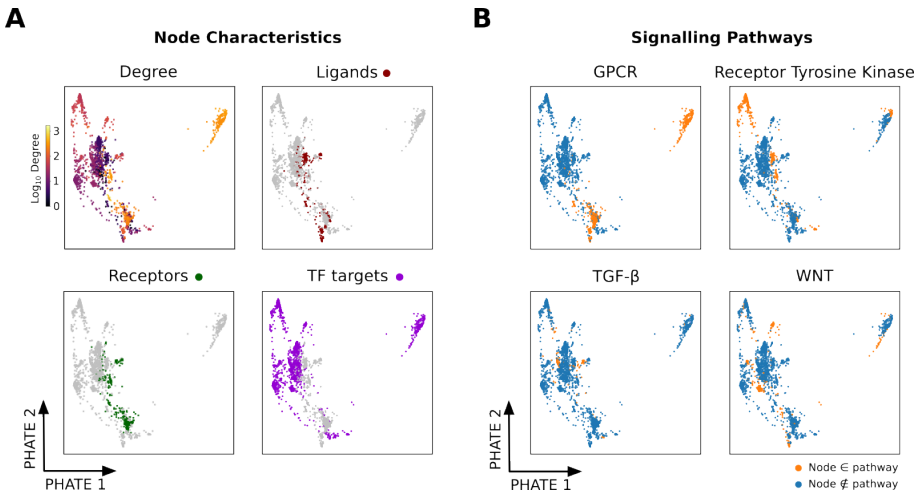


Figure 6.3: Information Preservation in Low-Dimensional KG Embeddings. A) PHATE of embedded KG nodes coloured by node-intrinsic properties. **B)** PHATE of embedded KG nodes coloured by relational signalling annotations. GPCR, G protein-coupled receptors.

To capture the complex relational information in a simpler format amenable to downstream analyses, directed heterogeneous knowledge graphs can be embedded into low dimensional tabular representations. Methods like the classical TransE [182] and its derivatives, graph convolutional networks, and hyperbolic embeddings [199], represent some of different approaches to learn the structure of KGs.

I used the TransR method [180] to embed the LRT-KG into a 50-dimensional space (Chapter 2), whose PHATE representation suggests that the embedding method captures topological differences between the distinct node types in the graph (Figure 6.3A). Node degree also seems to drive some of the topology in the PHATE representation of the embedding, and it would appear that TF targets are the most promiscuous nodes with higher degrees, followed by receptor nodes and finally ligands (Figure 6.3A). However, care must be taken when making these comparisons for three node classes are imbalanced.

Functional biological information encoded by the edges seems to also be captured in the embedded graph. Signalling pathways belonging to the Signal Trans-

duction category in Reactome, which should cover all three types of node in the LRT-KG, were mapped to gene node embedding. The resulting distribution of pathways, occupying discrete and specific regions of PHATE representation (Figure 6.3, appears to suggest that relational information from the KG is also conserved in the 50-dimensional embedding.

6.4 Projecting Cells as Signals on the KG

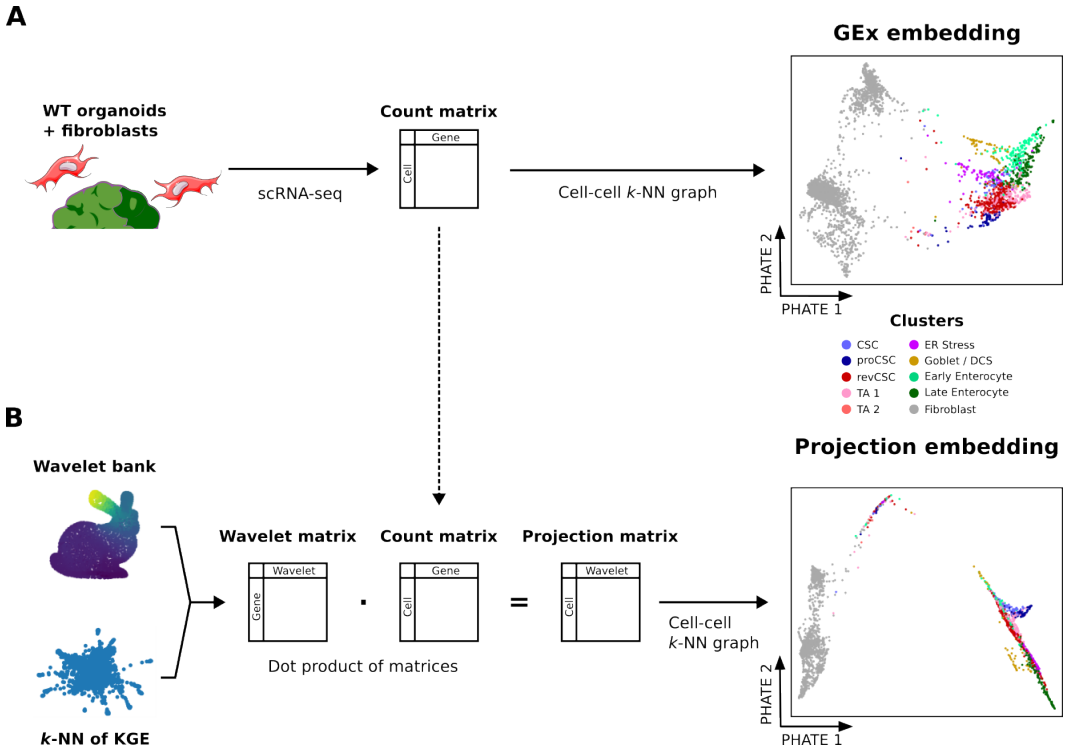


Figure 6.4: Projection of GEx Profiles on the LRT-KG. A) scRNA-seq datasets of WT organoid and fibroblast co-cultures are used for the projection. **B)** Wavelet diffusion is applied to the LRT-KG to generate a *nodeXwavelets* matrix onto which the sequencing data is projected. Colours on PHATE plots represent cell clusters.

Using a WT organoid and fibroblast co-culture scRNA-seq dataset from Chapter 4 (Figure 6.4A) I could explore the usefulness of the LRT-KG embedding to describe a cell's gene expression (GEx) profile as projected on a cell communication graph.

That particular dataset was employed because I had previously established, using cell-cell communication analysis tools and subsequent MC validation by Dr. Xiao Qin [1] (see Chapter 4 for Figure 4.6A), that the fibroblast cells engage in active communication with the organoid cells, in particular toward the revCSC state and adjacent areas of the colonic stem compartment.

When the transcriptomic data is used to generate a PHATE embedding the two distinct cell types are easily resolved, and so are the heterogeneous cell states within the colonic organoid epithelia (Figure 6.4A).

To project these cellular GEx profiles on the LRT-KG I first applied a diffusion

wavelet transform to a k-NN representation of the LRT-KG embedding, thus generating a $nodeXwavelets$ matrix where the first axis corresponds to the gene nodes of the LRT-KG (Figure 6.4B).

Leveraging the shared feature axis between the $nodeXwavelets$ and scRNA-seq $cellXgene$, I used the dot product ($\cdot$) operation to project the transcriptomic data as a $cellXwavelets$ matrix representation (Figure 6.4B). The projected data can be treated as the scRNA-seq count matrix from above to compute cell-cell k-NN graphs and two-dimensional embeddings.

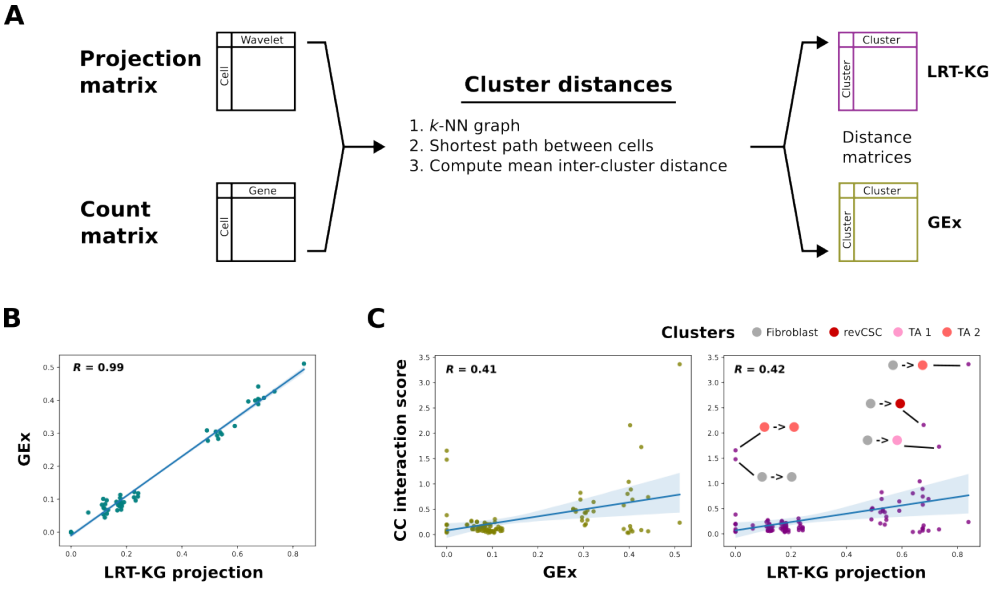
The resulting projection seems to non-quantitatively resemble the GEx profile on a PHATE space, wherein cell type is easily resolved. There appears however that there is some signal loss during the projection process, for epithelial heterogeneity is reduced (Figure 6.4B).

To quantitatively asses the projection results I not only compared it with the GEx data but also with the interaction strength predictions between cluster pairs in the data (see Chapter 2 for more details).

Average distances between cluster pairs in the LRT-KG projected space and the GEx space were computed based on their k-NN representations (Figure 6.5A) and found to be highly correlated (Figure 6.5B). A weak positive correlation ($R = 0.42$) between interacting cluster pairs and their distances was observed both in the GEx and highly similar projected spaces (Figure 6.5C).

Finally, the inter-cluster distance matrices (Sup. Tables C.11 and C.12) were scaled and subtracted to compare the differences between the GEx and projected profiles. Results revealed no distance shortening after projection between the highly interacting fibroblast and revCSC or TA clusters. Instead, projection lowered relative distances around the secretory cells and magnifying distances between the TA and ER stress states (Figure 6.5D).

In summary these results suggest an insufficient diffusion step prior to data projection, as shown by the similarities between the GEx and projected spaces, and a small degree of signal loss, eroding some the transcriptomic signal unique to secretory cells.



D

	CSC	proCSC	revCSC	TA1	TA2	ER Stress	Goblet / DCS	Early Enterocyte	Late Enterocyte	Fibroblast
CSC	0.000	0.007	0.038	0.058	0.048	0.064	-0.036	0.050	0.027	0.009
proCSC	0.007	0.000	0.015	0.034	0.025	0.041	-0.059	0.027	0.004	-0.015
revCSC	0.038	0.015	0.000	0.064	0.053	0.069	-0.030	0.056	0.033	0.020
TA1	0.058	0.034	0.064	0.000	0.073	0.090	-0.010	0.076	0.053	0.038
TA2	0.048	0.025	0.053	0.073	0.000	0.078	-0.021	0.067	0.045	0.000
ER Stress	0.064	0.041	0.069	0.090	0.078	0.000	-0.004	0.081	0.058	0.044
Goblet / DCS	-0.036	-0.059	-0.030	-0.010	-0.021	-0.004	0.000	-0.020	-0.043	-0.061
Early Enterocyte	0.050	0.027	0.056	0.076	0.067	0.081	-0.020	0.000	0.044	0.031
Late Enterocyte	0.027	0.004	0.033	0.053	0.045	0.058	-0.043	0.044	0.000	0.013
Fibroblast	0.009	-0.015	0.020	0.038	0.000	0.044	-0.061	0.031	0.013	0.000

Figure 6.5: Comparison of GEx and LRT-KG Projected Profiles. A) Inter-cluster distances are computed on GEx and projected spaces. **B)** Correlation between the two distance spaces. **C)** Correlation between cell-cell communication interaction scores and the distance spaces. Colour annotations reflect highly interacting cluster pairs. **D)** Scaled differences between the two cluster spaces. Cells are coloured according to the distance difference between a pair of cluster. R, Pearson correlation score.

6.5 Conclusions

In this chapter I have assembled a knowledge graph for cell communication that captures relational information between ligands, receptors and downstream targets of transcriptional factors. The assembled LRT-KG is comparable in size and graph characteristics to the curated OmniPath database, albeit with a lower number of nodes but enhanced hierarchical structure due to the reductionist approach of limiting signalling flow into a single direction from secreting to receiving cells and the latter's intra-cellular responses.

From this complex heterogeneous directed LRT-KG, methods like TransR can

learn a lower-dimensional embedding that captures the original node characteristics of the graph and even biological information in the form of signalling pathways encoded in the relations between nodes. The resulting LRT-KG embedding is a relatively simple tabular representation of the cellular communications LRT-KG onto which we can project the transcriptomic profile of cells via wavelet diffusion.

Projection results revealed similar PHATE embeddings and high inter-cluster distance correlation between the gene expression and projection spaces, suggesting that the diffusion process within the graph is not of a sufficient degree and remains too reliant on the graph's nodes rather than on its structure. While the similarities with the GEx profile do validate the projection approach, and some degree of correlation between both spaces was expected, the lacking diffusion step results in the projected space being unable to differentially capture inter-cellular communications between the interacting stromal and epithelial cells of WT organoid and fibroblast co-cultures.

Chapter 7

Discussion and Future Perspectives

7.1 Building Accessible and Automated Tools for MC Data Analysis

In this work I have shown CyGNAL's capabilities, describing in detail its design and inner mechanisms, and outlining its usefulness with regards to the analysis of MC datasets.

The main testament for the usefulness of the tool is the fact that it has become a part of routine MC analyses in our lab. With its support for plain text to FCS inter-compatibility (Chapter 2 and Figure 3.1), users can seamlessly integrate with MC platforms such as Cytobank. Given that the user only needs to run simple Python commands on the terminal to use CyGNAL, it has been readily adopted in day-to-day lab use even by users with no advanced computing experience. As I have shown in Chapters 2 and 3, CyGNAL is able to perform a comprehensive analysis of changes occurring across multiple conditions of the often wide MC experimental systems. Designed for the study of PTM signalling changes, CyGNAL's computation of EMD and DREMI scores resolves marker intensity and connectivity changes (Figure 3.2). The intuitive and customisable interactive Shiny-Apps allow for exploratory and close to publication-grade visualisation of the results (Figure 3.3). Tools are meant to be used, and that publications by colleagues such as Michelozzi *et al*. [186] employed CyGNAL is a testament to its relevance.

Originally meant as a simple exercise in curiosity-driven exploration after notic-

ing the correlation between so called PTM and 'cell-state' markers, and empowered by the tediousness of manually gating the datasets in our lab, the RF cell-state classifier has become a convenient tool to automate cell-state labelling of MC datasets in relation to cell-cycle phases.

Built around a simple Random Forest (RF) architecture, the RF classifier benefits from the fundamental gate-like logic of both decision trees and the manual cell-state gating process (Figure 3.4). However, I expect the classifier to suffer from generalisation issues when dealing with external data labelled using different workflows. Furthermore, even if it leverages fuzzy logic to match channel names from the model to the input data, the classifier still relies on matching markers found in both the training and test datasets. While the markers dedicated to apoptosis and cell-cycle phases generally belong to the less variable portions of MC panel design (Table 2.2), this can still pose an inconvenience when deploying the model. However, I have also shown how weak points such as low performance for apoptotic class prediction using the 5-marker MC model (Figure 3.5), can be effectively addressed by just the addition of an additional apoptotic marker to the panel design (Figure 3.5A). Furthermore, the model seems resilient to cell-type composition and even to broad cell-state changes induced by chemotherapy (Figure 3.5B-C).

Hearkening back to the link between PTMs and cell-cycle, the 10-marker MC model also reveals how certain PTMs prove more informative when training than *bona fide* 'cell-state' markers (Figure 2.2E). Furthermore, discrepancies between expected cell-state and PTM correlations from the literature and feature importance rankings have anecdotally been used to validate under-performing antibodies with high unspecific background staining.

Both these tools remain under continuous support, and I aim to eventually merge both code bases and integrate automated cell-state classification into CyGNAL using pre-built classifier models or allowing for the generation of new models based on specific user-provided labelled data. CyGNAL could also be augmented by the addition of PHATE [92] as an alternative DR step, implementing a new Shiny-App to visualise the embeddings and overlay user-selected metadata or antibody intensities.

7.2 Charting Stromal and Oncogenic Regulation of CSC Polarisation

Single-cell technologies can describe cell-cell communications and cell-type transitions in complex organoid settings and *in vivo* tissues [45, 49, 29]. As shown in Qin *et al.* [4], a heterocellular colonic epithelia organoid system can be employed in experimental designs covering the effects of both intrinsic CRC oncogenic mutations and extrinsic environmental cues. However, the directed and limited nature of the MC antibody panels used in Qin *et al.* [4] presented with a limiting factor towards a detailed description of colonic organoid epithelial polarisation by intrinsic and extrinsic cues.

Therefore, in Chapters 4 and 5 I have employed a multiplexed scRNA-seq analysis of heterocellular CRC organoid cultures (Figure 4.1) to chart a continuous landscape of intrinsic and extrinsic regulation of CSC states. I have found that stromal cues transition the epithelia towards the revCSC state, oncogenic signalling pushes the organoid towards proCSC, and exogenous ligands overlapping with both stromal and oncogenic signalling cues can polarise towards both states at once (Figure 4.3). I have also developed a method to capture these transitional processes, the valley-ridge (VR) score (Figure 5.1), and established a workflow to project it onto Waddington-like data-driven landscapes (Figure 2.3). The work presented in this thesis was paired with complementary MC experiments in Qin & Cardoso Rodriguez *et al.* [1], where we interrogated colonic stem cell regulation at scale to functionally understand the polarisation mechanisms (Appendix D).

First, I have shown that transcriptomic profiles of epithelial, fibroblast and macrophage cells from the heterocellular cultures can be used to describe inter-type heterogeneity and recapitulate the distinct epithelial compartments (Figures 4.1 & 4.2). The observed *Cd34* high and low fibroblast populations are reminiscent of *in situ* intestinal fibroblast heterogeneity, wherein *Cd34* expressing fibroblast from the bottom of the crypts support the intestinal stem niche [200] whereas *Cd34* low fibroblast are found above the crypt's bottoms and help maintain the BMP gradient needed for epithelial differentiation [192]. While I observed some transcriptional differences

between these two fibroblast populations (Sup. Figure B.1), their regulation of the epithelial compartment remained consistent (Sup. Figure B.2), possibly due to shared secreted signalling between the two. Myeloid macrophage transcriptomes formed a continuum trajectory of putative inflammation-related roles (Sup. Figure B.3), unlike the distinct fibroblast and epithelial populations. However, neither macrophages as a whole nor the extremes of their transcriptional continuum differentially regulated the epithelial cells.

The healthy small intestinal and colonic epithelia is supported by a stem cell niche at the bottom of the crypts regulated by both intrinsic and stroma-secreted signalling gradients. These traditional colonic stem cell (CSC) however, are not the sole stem cell state, with less common low-proliferative revival CSC (revCSC) being able to replenish the CSC niche and repair the epithelial tissue in response to tissue damage [30]. Here I have shown how these revCSC are enriched by stromal WNT and TGF-β when WT organoids are co-cultured with fibroblasts (Figures 4.3A & 4.6B), and how revCSC also resemble public descriptions of the same population and a "foetal"-like state [158] (Figure 4.7A).

The gradient of organoids with accumulating oncogenic mutations revealed how a hyper-proliferative CSC (proCSC) state is enriched in CRC organoids (Figure 4.3B). These cells are present in lower numbers in WT and *shApc* organoids, but quickly dominate the landscape of stunted absorptive and secretory differentiation in the *shApc* and $Kras^{G12D/+}$ (AK), and *shApc*, $Kras^{G12D/+}$ and $Trp53^{R172H/-}$ (AKP) colonic organoids (Figure 4.3C). proCSC were found to be transcriptionally similar to other cells from mouse models and human CRC (Figure 4.7A).

With a clear differential regulation by extrinsic stromal cues and intrinsic oncogenic signalling, polarisation of WT colonic epithelia towards both proCSC and revCSC could nonetheless be achieved via exogenous WENR added to the culture media (Figure 4.3C). These findings, together with subsequent MC validation [1] of the signalling hubs identified via cell-cell communication analysis, suggest that both states are part of a shared polarisation landscape with overlapping signalling pathways that compete to establish colonic epithelial cell-fate. In this context, the

observed breakdown of fibroblast-to epithelia communications in CRC organoids (at least partly due to downregulation of key signalling receptors by the epithelial cells) seems to suggest that intrinsic oncogenic cues dominate extrinsic stromal cues (Figure 4.6). The interplay between the two with regard to proCSC and revCSC polarisation is explored further in Qin & Cardoso Rodriguez *et al.* [1], where we established that TGF-β can induce revCSC-like cells in CRC organoids in the context of low PI3K signalling, supporting the suggested role of revCSC as a drug-resistant state in CRC that can drive relapse after chemotherapy [32, 3].

In silico analysis of cellular dynamics identifies revCSC as a terminal cell-fate (Figure 4.5E), suggesting that polarisation of the colonic epithelia towards revCSC is achieved via plasticity-driven transitional processes from adjacent cell-states. In contrast, proCSC is consistently identified as an initial population (Figure 4.5) whose dominance of the epithelia seems to be achieved due to its high proliferative potential.

Therefore, I postulated that cellular pluripotency scores and rates of transcriptomic change could capture the cellular dynamics of such systems, providing for an avenue towards generation of data-driven Waddington-like landscapes of cellular differentiation and plasticity. The valley-ridge (VR) score described in Chapter 5 synthesises both CCAT and RNA velocity vector length metrics to capture coarse pluripotency changes and global transcriptomic structure with PHATE. Finer details at a local level capture the availability of cell-states as determined by RNA velocity (Figure 5.1). The methodology presented also incorporates with a landscape projection pipeline (Figure 2.3). The VR landscapes reconstruct the shared landscape of colonic stem cell polarisation, presenting revCSC as an accessible epithelial fate in the presence of stromal ligands, whereas intrinsic oncogenic signalling trap the organoid in a highly pluripotent yet isolated proCSC fate, refractory to stromal signals that otherwise would polarise the cells towards revCSC (Figure 5.2).

The work presented in these two chapters presents with some notable limitations, such as a lack of non-organoid *in situ* validation: with the only effort towards validating the findings being achieved via *in silico* signature matching and data integration (Figure 4.7). Furthermore, non-paracrine stromal regulation, specially

given the well-known role of fibroblasts as extra-cellular matrix re-modellers, has not been deeply explored in this study. It is also worth noting that a line of normal murine intestinal fibroblasts was used in the organoid co-cultures, rather than pairing the CRC organoids with cancer-associated fibroblasts. This later point will be addressed in subsequent studies at the lab by attempting to match patient-derived organoids with cancer-associated fibroblasts from the same donor. Further work regarding the cross-validation with human data of the proCSC and revCSC cell identities and functional characteristics is being carried out as part of the peer-review process of the work presented in Qin & Cardoso Rodriguez *et al.* [1]. Furthermore, additional improvements to the VR score and landscape generation will be implemented during the later stages of my project. Aiming to increase the tool's accessibility and ease of use, the current Jupyter Notebook format will be adapted to the nbdev framework (`https://nbdev.fast.ai/`). VR landscapes will be packaged as a tool, *VR Land* (`github.com/FerranC96/VRland`), which will be distributed as an interactive web-app to facilitate the exploration of the 3-dimensional landscapes generated.

In conclusion, these results describe fibroblasts as key stromal regulators of the colonic stem compartment, orientating epithelial stem cell fate via secreted WNT and TGF-β . Stromal regulation competes with, and is ultimately trumped by, the proCSC-enriching organoid-intrinsic oncogenic cues. Further understanding concerning the regulation of proCSC and revCSC fates might suggest new avenues for cancer therapies. Indeed, given that revCSC has already been described as a drug-tolerant persister state [32], blocking the plastic processes controlling its accessibility might be a valid strategy to limit the emergence of chemotherapy resistance.

7.3 Knowledge Graphs for Cell Communication

While cellular communications are commonly understood to be a complex process both at the inter- and intra-cellular levels, there is a lack of tools aiming to capture the causal and directed nature of the process. Coupled with emerging multi-modal approaches that could measure gene and protein expression, including PTMs, methods capturing both paracrine secreted signalling and cell-state responses to extrinsic cues should describe a holistic view of cellular communications.

In Chapter 6 I have assembled a directed and hierarchical ligand-receptor-target KG (LRT-KG) from publicly available databases that aims to capture the cellular signalling occurring both between interacting cells and within a cell receiving extrinsic cues (Figure 6.1). Aiming to apply this new method to study cell communications within the WT organoid and fibroblast co-culture in a holistic manner, the assembled KG has a complexity comparable to the curated OmniPath database (Figure 6.2, Table 2.3). Nonetheless, I have shown that knowledge graph embedding approaches can learn a simpler tabular representation of the KG that conserves the biological information encoded within it; including relational information between the gene nodes regarding pathway annotations (Figure 6.3).

Using a wavelet-based diffusion step and projecting the scRNA-seq organoid co-culture data (Figures 2.4 & 6.4), I have successfully shown that projected cellular profiles diffused on the KG preserve the information encoded in the original transcriptomic data representation. However, the projected profiles appear to be too similar to the gene expression (GEx) data (Figure 6.4). Indeed, when inter-cluster distances are computed, no significant change was detected between the original and projected views; with fibroblasts and revCSCs, found to be closely interacting by cell-cell communication, remaining at comparable proximity to their prior GEx profiles (Figure 6.5).

Most likely explained by an insufficient diffusion process, alternative approaches are being explored in conjunction with my collaborators at Yale University; such as the work on directed scattering transforms presented at the Graph Signal Processing Workshop 2023 (`https://ferranc96.github.io/posts/GSPw23/`).

With a robust diffusion process, the method performance could also be benchmarked by leveraging spatial data as done by alternative cell-cell communication approaches [197]. Multi-modal data could also be projected on a modality-agnostic feature-feature KG with both protein and gene nodes. This approach should be able to more confidently call inter-cellular interactions via ligand-receptor expression, and intra-cellular responses via PTM profiles and expression of transcription factor targets.

In summary, a balance between limiting signal loss (determined by the nodes in the KG) and adequate diffusion approaches (ensuring sufficient information on the graph structure itself is captured during data projection) is necessary for such a holistic cell communication method to perform adequately. It would appear then, that the current implementation requires of further work on the later point. Published approaches exist to tackle similar problems [132, 198], but the aim of treating the cells as signals to be compared on a gene-gene graph (or other *omic* features), remains to my knowledge unique to the efforts presented here and worth pursuing specially considering multi-modal profiles could be projected on such feature-feature KGs.

Appendix A

pyKrack

A.1 Introduction

Biological signalling can be modelled as a directed network, where nodes represent genes/proteins and edges represent signalling interactions.

The hierarchy of such a network can be quantified using various metrics, including the Krackhardt hierarchy score. This score measures the degree to which the network exhibits a perfect hierarchy, with higher scores indicating a greater hierarchy.

In R the sna package presents methods to compute graph hierarchy including Krackhardt's score, and there are other hierarchy scores implemented in Python such as Flow Hierarchy Score [201]. However, despite its utility, there is currently no native implementation of the Krackhardt hierarchy score in Python.

A.1.1 Krackhardt Hierarchy Score

The Krackhard hierarchy score was introduced by David Krackhardt [Krackhardt, David. (1994). Graph Theoretical Dimensions of Informal Organization. Computational Organization Theory. 89], where he defined it as:

> *The graph hierarchy condition states that in a digraph D, for each pair of points where one (Pi) can reach another (Pj), the second (Pj) can't reach the first (Pi). For example, in a formal organization chart a high level employee can reach through the chain of command her subordinate's subordinate. If the formal organization is working "properly", this lower*

level employee can't simultaneously reach the high level employee. To measure the degree of hierarchy of digraph D, a new digraph Dr must be created. Dr is defined as the reachability digraph of D. Each point in D exists in Dr; moreover, the line (Pi,Pj) exists in Dr if and only if Pi can reach Pj in D. If D is graph hierarchic, then Dr will have no symmetric lines in it (i.e. if the line (Pi,Pj) exists in Dr then the line (Pj,Pi) does not).

The degree of hierarchy then is defined as

$$GraphHierarchy = 1 - [V/MaxV]$$

where V is the number of unordered pairs of points in Dr that are symmetrically linked and $MaxV$ the number of unordered pairs of points in Dr where Pi is linked to Pj or viceversa.

A.2 Hierarchy Computation

Based on the definition above, I wrote a small Python package to compute the Krackhardt hierarchy score. Built around a main function that computes the hierarchy score (Listing A.2), the pykrack package (`ferranc96.github.io/pyKrack`) also includes a helper function to describe general properties of a directed graph and computes an alternative hierarchy score.

```
1
2  def compute_hierarchy(G, metric="pykrack"):
3      """
4      Compute one of the possible hierarchy scores
5
6      Parameters
7      ----------
8      G
9          Directed NetworkX graph
10     metric : str
11         Type of hierarchy metric to compute. Accepted types are:
12         'pykrack' for this module's implementation of the Krackhardt score.
13         'rsnakrack' for the sna implementation in R.
14         'hierarchy_flow' for the Luo and Magee 2011 as implemented in the
       NetworkX package.
```

```
15
16      Returns
17      -------
18      score : float
19          One of the possible hierarchy scores
20      """
21
22      #Ensure Graph is DirectedGraph
23      if not G.is_directed():
24          raise Exception
25      #Ensure Graph is of DiGraph() format
26      G = nx.DiGraph(G)
27
28      if metric == "pykrack": #Python implementation
29          #Compute transitive closure of graph to get the reachability graph
30              #[contains an edge (i,j) if there is a path from i to j in the
    original graph]
31          acyclic = 0
32          try:
33              nx.find_cycle(G)
34          except:
35              print("Acyclic graph")
36              acyclic = 1
37          if acyclic == 1:
38              Gr = nx.transitive_closure_dag(G)
39          else:
40              Gr = nx.transitive_closure(G, reflexive=None)
41          symmetric_dyads = 0
42          non_null_dyads = 0
43          n = len(Gr.nodes())
44          #Count the number of non-null symmetric dyads
45          for pair in product(Gr.nodes(), Gr.nodes()):
46              if Gr.has_edge(pair[0],pair[1]) or Gr.has_edge(pair[1],pair[0]): #
    Non-null dyad
47                  non_null_dyads+=1
48                  if  Gr.has_edge(pair[0],pair[1]) == Gr.has_edge(pair[1],pair
    [0]): #Symmetric!
49                      symmetric_dyads+=1
50          #Raise exception if graph has no edges!
51          if non_null_dyads == 0:
52              raise Exception
53          score = 1 - (symmetric_dyads / non_null_dyads)
54
55      elif metric == "rsnakrack": #R implementation from the sna package
56          try:
57              base = importr("base")
```

```
58              sna = importr("sna")
59              score = sna.hierarchy(nx.to_numpy_array(G), measure="krackhardt")
    [0]
60          except:
61              print("R package sna was not found. Please install manually!")
62              print("Computing hierarchy flow instead")
63              snafail_flag = 1
64              score = nx.flow_hierarchy(G)
65
66      elif metric == "hierarchy_flow": #Networkx's hierarchy flow implementation
67          score = nx.flow_hierarchy(G)
68
69      # elif metric == "all": #This will eventually return a dict with all
        metrics
70
71      else: # metric argument broken
72          raise Exception
73
74
75      return score
```

Listing A.1: Main pyKrack function. The *compute_hierarchy* function takes in a directed graph and computes its hierarchy flow score or the Krackhardt hierarchy score using an existing R implementation or a novel one in Python.

A.3 Notebook-Centric Implementation

This package has been implemented using the nbdev framework (`nbdev.fast.ai`). This technology allows for a notebook-centric approach to software development and distribution, including automation of documentation sites to continuous integration actions that automate package releases.

I have leveraged nbdev to publish this tool as a package in Pypi (`pypi.org/project/pykrack`), and as a technology demonstrator for an upcoming deployment of the VR score landscapes.

Appendix B

Supplementary Figures

B.1 Figures related to Chapter 4

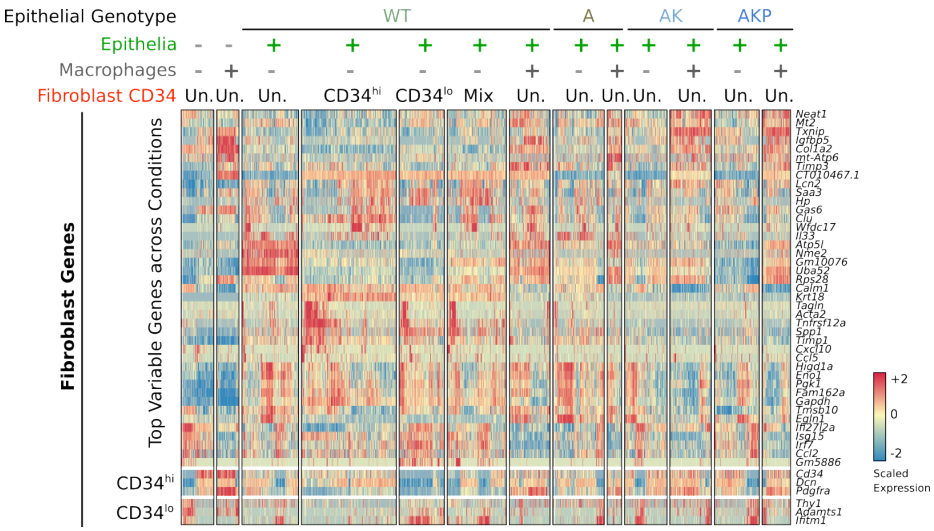


Figure B.1: Fibroblast DE Analysis. Differential gene expression analysis of fibroblasts regulated by epithelial organoids and macrophages.

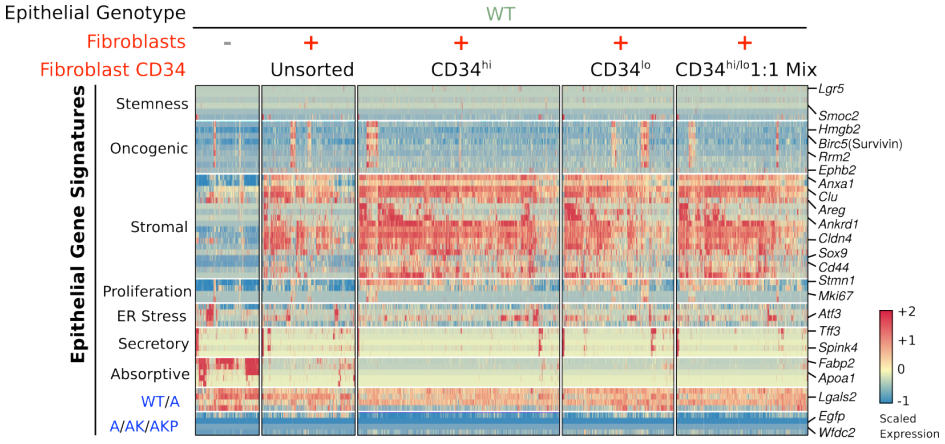


Figure B.2: Epithelial DE Analysis by Fibroblast-Subtype. Differential gene expression analysis of WT colonic organoids co-cultured with unsorted, CD34^{hi}, CD34^{lo}, and a 1:1 mix of CD34^{hi}:CD34^{hi} colonic fibroblasts.

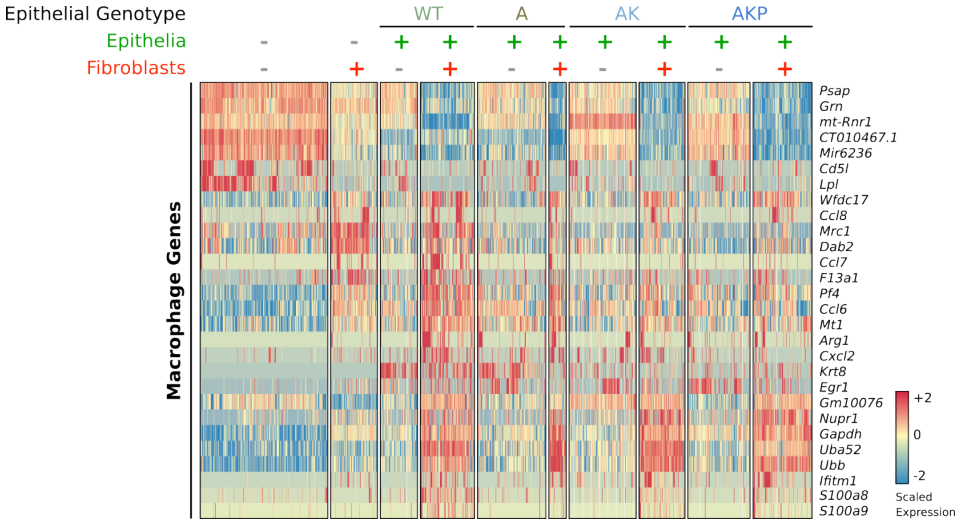


Figure B.3: Macrophages DE Analysis. Differential gene expression analysis of macrophages regulated by epithelial organoids and fibroblasts

Appendix C

Supplementary Tables

C.1 Gene Data

Table C.1: Colonic Epithelia Gene Markers (1/2). Markers of epithelial populations and organoid genotypes. Derived from literature and DE analysis of our data.

Gene	Annotation
Lgr5	CSC
Pla2g2a	CSC
Lrig1	CSC
Smoc2	CSC
Aqp5	CSC
Agr2	CSC
Cenpa	proCSC
Hmgb2	proCSC
Birc5	proCSC
Tuba1b	proCSC
Ube2c	proCSC
Rrm2	proCSC
Hells	proCSC
Cdk1	proCSC
Ephb2	proCSC
Anxa1	revCSC
Ly6a	revCSC
Clu	revCSC
Basp1	revCSC
Areg	revCSC
Ccn1	revCSC
Ccn2	revCSC
Ankrd1	revCSC
Ctla2a	revCSC
Pmepa1	revCSC
Marcksl1	revCSC
Cldn4	revCSC
F3	revCSC

Table C.2: Colonic Epithelia Gene Markers (2/2).

Gene	Annotation
Ecm1	revCSC
Sox9	revCSC
Cd44	revCSC
Itga2	revCSC
Fn1	revCSC
Stmn1	Replication
Ccnd1	Replication
Mki67	Replication
Ccnb1	Replication
Hspa5	ER stress
Ddit3	ER stress
Atf3	ER stress
Atf4	ER stress
Tff3	Secretory
Atoh1	Secretory
Muc2	Secretory
Spink4	Secretory
Reg4	Secretory
Fabp2	Absorptive
Aldob	Absorptive
Apoa1	Absorptive
Fabp1	Absorptive
Apoa4	Absorptive
Msln	WT˙A
Lgals2	WT˙A
Rps4l	WT˙A
Gsta3	WT˙A
Hopx	A˙AK˙AKP
pEGFP	A˙AK˙AKP
Wfdc2	A˙AK˙AKP
Ly6c1	A˙AK˙AKP

Table C.3: Cell-Cycle Gene Lists (1/6). Table of cell-cycle genes adapted from Tirosh *et al.* [8] and Macosko *et al.* [9], the former using a human melanoma cell line and the later both human and mouse models to link gene expression with cell cycle phases. The original tables provided in the publication were pooled together, duplicated genes were dropped, and human symbols were translated to mouse using BioMart. Finally, genes whose expression could not be detected in any of the mouse organoid experiments were dropped from the list. The resulting table contains 98 genes associated with S-phase, 248 with both G2 and M-phase, and 202 with G1.

S-phase	G2 & M-phase	G1
Abcc5	Ahi1	1700009N14Rik
Abhd10	Akirin2	2700049A03Rik
Asf1b	Ankrd40	Acd
ATAD2	Anln	Acyp1
Bbs2	Anp32b	Adamts1
Bivm	Anp32e	Agfg1
Blm	Ap3d1	Agpat3
Bmi1	Arhgap19	Ak6
Brca1	Arl4a	Akap13
Brip1	Arl6ip1	Amd1
Cald1	Armc1	Amd2
Calm2	Asxl1	Ankrd10
Casp2	Atf7ip	Anp32e
Ccdc14	Atl2	Antxr1
Ccdc150	Aurka	Apex2
Ccdc84	Aurkb	Arglu1
Cdc45	Bclaf1	Bag3
Cdc7	Birc2	Bard1
Cdca5	Birc5	BC048507
Cdkn2aip	Bora	Brd7
Cenpm	Brd8	Btbd3
Cenpq	Bub1	Capn7
Cers6	Bub3	Casp2
Chml	Cadm1	Casp8ap2
Coq9	Casp3	Cbx3
Cpne8	Cbx5	Ccne1
Crebzf	Ccdc107	Ccne2
Crls1	Ccdc88a	Cdc25a
Depdc7	Ccdc90b	Cdc42
Dhfr	Ccna2	Cdc6
Dna2	Ccnb2	Cdca7
Dnajb4	Ccnf	Cdca7l
Donson	Cdc16	Cdk7
Dscc1	Cdc20	Cdkn3
Dync1li2	Cdc25b	Cep57
E2f8	Cdc25c	Cep70
Eif4ebp2	Cdc27	Chaf1a

Table C.4: Cell-Cycle gene lists (2/6).

S-phase	G2 & M-phase	G1
Ercc5	Cdc42ep1	Chaf1b
Esco2	Cdca2	Clspn
Exo1	Cdca3	Cnih4
Ezh2	Cdca8	Crebzf
Fanca	Cdk1	Ctr9
Fanci	Cdkn1b	Ctsd
Fen1	Cdkn2c	Cwc15
Gclm	Cdr2	Dcp1a
Gm13547	Cenpa	Dctn6
Hells	Cenpe	Dexi
Ints7	Cenpf	Dis3
Kat2a	Cenpl	Dkc1
Kat2b	Cep350	Dnajb6
Lmo4	Cep55	Dnajc3
Lyrm7	Cfd	Donson
Man1a2	Cflar	Dscc1
Map3k2	Chek2	Dsp
Mastl	Cit	Dtl
Mbd4	Ckap2	Dynll1
Mcm8	Ckap2l	E2f1
MLF1IP	Ckap5	Eif2a
Mycbp2	Cks1b	Eif4e
Nab1	Cks2	Elp3
Nfe2l2	Cnot10	Esd
Nsun3	Cntrob	Fam122a
Nt5dc1	Ctcf	Fam189b
Nup160	Ctnna1	Flad1
Ogt	Ctnnd1	Fopnl
Orc3	Cyth2	Foxk2
Osgin2	Dcaf7	Fxr1
Phip	Depdc1a	G3bp1
Phtf1	Depdc1b	Gata2
Phtf2	Dhx8	Gins2
Pkmyt1	Diaph3	Gins3
Pola1	Dlgap5	Gm12666
Prim1	Dnajb1	Gm45713
Ptar1	Dr1	Gm49369
Rad18	Dzip3	Gmnn
Rad51	E2f5	Gnb1
Rad51ap1	Ect2	Grpel1
Rbbp8	Entpd5	Gspt1
Reep1	Espl1	Gtf3c4
Rfc2	Fadd	Hells

Table C.5: Cell-Cycle gene lists (3/6).

S-phase	G2 & M-phase	G1
Rhobtb3	Fam83d	Hif1a
Rmi1	Fan1	Hmg20b
Rpa2	Fancd2	Hmgcr
Rrm1	Foxm1	Hoxb4
Rrm2	Fyn	Hras
Rsrc2	G2e3	Hsd17b11
Sap30bp	Gabpb1	Hsf2
Slc38a2	Gadd45a	Hspa8
Sp1	Gas1	Ilf2
Srsf5	Gas2l3	Insr
Svip	Gm10184	Ints8
Top2a	Gm28635	Ivns1abp
Ttll7	Got1	Jmjd1c
Tyms	Grk6	Kdm5b
Ube2t	Gtse1	Kif5b
Ubl3	Haus8	Kpnb1
Usp1	Hcfc1	Kras
Zwint	Hint3	Larp1
	Hipk2	Larp7
	Hjurp	Lnpep
	Hmg20b	Lrif1
	Hmgb2	Luc7l3
	Hmgb3	Lyar
	Hmmr	Mcm2
	Hp1bp3	Mcm4
	Hps4	Mcm5
	Hs2st1	Mcm6
	Hspa13	Mdm1
	Hspa8	Med31
	Ifnar1	Morf4l2
	Iqgap3	Mri1
	Katna1	Mrpl19
	Kctd9	Mrps18b
	Kdm4a	Mrps2
	Kif11	Msh2
	Kif14	Msl1
	Kif20b	Mtpn
	Kif22	Nasp
	Kif23	Ncoa3
	Kif2c	Nfia
	Kif5b	Nfic
	Kifc1	Nktr
	Kifc5b	Npat

Table C.6: Cell-Cycle gene lists (4/6).

S-phase	G2 & M-phase	G1
	Klf6	Nucks1
	Klf9	Nufip2
	Kpna2	Nup37
	Lbr	Nup43
	Lix1l	Odf2
	Lmna	Opn3
	Lmnb1	Orc1
	Mad2l1	Osbpl6
	Mcm4	Pak1ip1
	Mdc1	Pank2
	Melk	Pbk
	Mgat2	Pcdh7
	Mid1	Pcf11
	Mis18bp1	Pcna
	Mki67	Plcxd1
	Mnd1	Plin3
	Mzt1	Pms1
	Ncapd2	Pnn
	Ncapd3	POLD3
	Ncaph	Ppp2ca
	Ncoa5	Ppp2r2a
	Ndc80	Ppp6r3
	Neil3	Prc1
	Nek2	Psen1
	Nfic	Pttg1
	Nipbl	Rab23
	Nmb	Rad21
	Nr3c1	Recql4
	Nucks1	Rheb
	Nuf2	Rmi2
	Numa1	Rnf113a1
	Nup35	Rnf113a2
	Nup98	Rnpc3
	Nusap1	Rpl13a
	Odf2	Sec62
	Pbk	Skp2
	Pcf11	Slbp
	Pif1	Slc25a36
	Pknox1	Slc39a10
	Plk1	Snupn
	Poc1a	Srsf3
	Polq	Srsf7
	Pom121	Ssr3

Table C.7: Cell-Cycle gene lists (5/6).

S-phase	G2 & M-phase	G1
	Ppp1r10	Stag1
	Ppp1r2	Syncrip
	Prpsap1	Taf15
	Prr11	Taf9
	Psmd11	Tcerg1
	Psmg3	Tipin
	Psrc1	Tle3
	Ptp4a1	Tmem138
	Ptpn9	Tob2
	Pwp1	Top1
	Qrich1	Topbp1
	Rad51c	Tra2a
	Rangap1	Troap
	Rbm8a	Tsc22d1
	Rbm8a2	Ttc14
	Rcan1	Tulp4
	Rccd1	Ube2d3
	Rdh11	Ubr7
	Rere	Uhrf1
	Rnf126	Ung
	Rnf141	Usp53
	Rnps1	Vangl1
	Rrp1	Vcl
	Sap30	Vps72
	Sephs1	Wdr76
	Sfpq	Wipf2
	Shcbp1	Wwc1
	Ska3	Yy1
	Smarcb1	Zbtb7a
	Smarcd1	Zcchc10
	Smc4	Zfp24
	Spag5	Zfp281
	Sptbn1	Zfp367
	Srf	Zfp593
	Srsf3	Zmynd19
	Ss18	Zranb2
	Stat1	
	Stil	
	Stk17b	
	Suclg2	
	Tacc3	
	Tfap2a	
	Thrap3	

Table C.8: Cell-Cycle gene lists (6/6).

S-phase	G2 & M-phase	G1
	Timp1	
	Tle3	
	Tmem138	
	Tmpo	
	Tnpo1	
	Tnpo2	
	Tomm34	
	Top2a	
	Tpx2	
	Traip	
	Trim59	
	Trip13	
	Trmt2a	
	Tsg101	
	Tsn	
	Ttf2	
	Ttk	
	Tuba1a	
	Tubb2a	
	Tubb4b	
	Tubb5	
	Tubd1	
	Txndc9	
	Txnrd1	
	Uaca	
	Ube2c	
	Ube2d3	
	Usp13	
	Usp16	
	Vangl1	
	Vps25	
	Vta1	
	Wsb1	
	Ywhah	
	Zc3hc1	
	Zfp207	
	Zfx	
	Zmym1	
	Znhit2	

Table C.9: Literature Gene Signatures (1/2). Metadata for the literature gene signatures characterising the various stem cell states in intestinal and colon epithelia, as well as certain key signalling pathways.

Name	Genes	Context	Species	Reference
lgr52MEX3A	91	CSC	Human	Alvarez et al. 2022
lgr5MEX3A	6	CSC	Human	Alvarez et al. 2022
Ssc1AYYAZ19	50	CSC	Mouse	Ayyaz et al. 2019
Ssc2bAYYAZ19	50	CSC	Mouse	Ayyaz et al. 2019
StemBUES22	9	CSC	Mouse	Bues et al. 2022
ProgenstemDALERBA11	4	CSC	Human	Dalerba et al. 2011
IscGREGORIEFF15	9	CSC	Mouse	Gregorieff et al. 2015
IscHAN20	5	CSC	Mouse	Han et al. 2020
IscLI17	38	CSC	Human	Li et al. 2017
StemcorrLI17	7	CSC	Human	Li et al. 2017
StemtaLI17	36	CSC	Human	Li et al. 2017
hEphb2MERLOS11	28	CSC	Human	Merlos et al. 2011
hLgr5MERLOS11	50	CSC	Human	Merlos et al. 2011
IscMERLOS11	49	CSC	Mouse	Merlos et al. 2011
Lgr5MERLOS11	103	CSC	Mouse	Merlos et al. 2011
Lgr5MOURAO19	98	CSC	Mouse	Mourao et al. 2019
munozLGR5MEX3A	139	CSC	Mouse	Muñoz et al. 2012
StemtaPELKA21	79	CSC	Human	Pelka et al. 2021
StemtasecPELKA21	94	CSC	Human	Pelka et al. 2021
OWNdeclusstem	24	CSC	Mouse	Qin & Cardoso et al. 2023
OWNsigstem	6	CSC	Mouse	Qin & Cardoso et al. 2023
eecMEX3A	13	Other	Human	Alvarez et al. 2022
gobletMEX3A	6	Other	Human	Alvarez et al. 2022
labelMEX3A	58	Other	Human	Alvarez et al. 2022
mucsecMEX3A	13	Other	Human	Alvarez et al. 2022
panethMEX3A	12	Other	Human	Alvarez et al. 2022
secPROGMEX3A	11	Other	Human	Alvarez et al. 2022
Ssc2aAYYAZ19	50	Other	Mouse	Ayyaz et al. 2019
Ssc2AYYAZ19	50	Other	Mouse	Ayyaz et al. 2019
ImmatureDALERBA11	5	Other	Human	Dalerba et al. 2011
iCMS2	288	Other	Human	Joanito et al. 2022
iCMS	58	Other	Human	Joanito et al. 2022
cryptPROLIFMEX3A	269	proCSC	Human	Alvarez et al. 2022
ki67MEX3A	62	proCSC	Human	Alvarez et al. 2022
CancerDALERBA11	3	proCSC	Human	Dalerba et al. 2011

Table C.10: Literature gene signature (2/2).

Name	Genes	Context	Species	Reference
ProlifDALERBA11	3	proCSC	Human	Dalerba et al. 2011
TumourLI17	9	proCSC	Human	Li et al. 2017
hProlifMERLOS11	176	proCSC	Human	Merlos et al. 2011
ProlifMERLOS11	258	proCSC	Mouse	Merlos et al. 2011
StemtaprolifPELKA21	89	proCSC	Human	Pelka et al. 2021
OWNdecluststemO	24	proCSC	Mouse	Qin & Cardoso et al. 2023
OWNsigprolif	4	proCSC	Mouse	Qin & Cardoso et al. 2023
OWNsigstemO	9	proCSC	Mouse	Qin & Cardoso et al. 2023
mex3aMEX3A	83	revCSC	Human	Alvarez et al. 2022
revSCMEX3A	5	revCSC	Human	Alvarez et al. 2022
Ssc2cAYYAZ19	50	revCSC	Mouse	Ayyaz et al. 2019
mex3aBARRIGAMEX3A	93	revCSC	Mouse	Barriga et al. 2017
RsBUES22	6	revCSC	Mouse	Bues et al. 2022
epiHrCANELLAS22	92	revCSC	Human	Cañellas et al. 2022
RegenscgGIL22	265	revCSC	Mouse	Gil Vazquez et al. 2022
RepGREGORIEFF15	8	revCSC	Mouse	Gregorieff et al. 2015
FetalHAN20	5	revCSC	Mouse	Han et al. 2020
RevscHAN20	4	revCSC	Mouse	Han et al. 2020
mustataFETALMEX3A	56	revCSC	Mouse	Mustata et al. 2013
OWNdecluststemS	24	revCSC	Mouse	Qin & Cardoso et al. 2023
OWNsigstemS	18	revCSC	Mouse	Qin & Cardoso et al. 2023
FetalYUI18	1184	revCSC	Mouse	Yui et al. 2018
yapMEX3A	8	Signalling	Mouse	Alvarez et al. 2022
MapkGO	23	Signalling	Human	Gene Ontology term
Pi3kGO	37	Signalling	Human	Gene Ontology term
YapGREGORIEFF15	24	Signalling	Mouse	Gregorieff et al. 2015
KrasGSEA	45	Signalling	Human	GSEA
WntHAN20	7	Signalling	Mouse	Han et al. 2020
YapHAN20	6	Signalling	Mouse	Han et al. 2020
MapkKEGG	155	Signalling	Human	KEGG
TgfbKEGG	80	Signalling	Human	KEGG
TgfbLI17	33	Signalling	Human	Li et al. 2017
WntLI17	18	Signalling	Human	Li et al. 2017
WntMORRAL20	60	Signalling	Human	Morral et al. 2020
Notch1MOURAO19	289	Signalling	Mouse	Mourao et al. 2019
OWNwntreceptors	17	Signalling	Mouse	Qin & Cardoso et al. 2023
YapWANG18	21	Signalling	Human	Wang et al. 2018

C.2 Knowledge Graph Data

```
    def CalculateWavelets(self,use_reduced=False,J=-1,epsilon=1e-3):

        # assert(self.P)

        if J == -1:
            J = int(np.log(self.N))
        self.J = J

        I = np.eye(self.N)
        self.wavelets = [I]
        P_j = np.linalg.matrix_power(self.P,2)

        print("Calculating Wavelets Using J = " + str(J))

        if use_reduced:
            #assert(self.N < 3000)
            Psi_j_tilde = column_subset(I-P_j, epsilon=epsilon)
            self.wavelets += [Psi_j_tilde]
            for i in tqdm(range(2,J)):
                P_j_new = np.linalg.matrix_power(P_j,2)
                Psi_j = P_j - P_j_new
                P_j = P_j_new
                self.wavelets += [column_subset(Psi_j,1e-3)]
        else:
            self.wavelets += [I-P_j]
            for i in tqdm(range(2,J)):
                P_j_new = np.linalg.matrix_power(P_j,2)
                Psi_j = P_j - P_j_new
                P_j = P_j_new
                self.wavelets += [Psi_j]
```

Listing C.1: Wavelet module. Function to compute the wavelet difussion transform. Part of a broader script kindly provided by Aarthi Venkat from Prof. Smita Krishnaswamy's lab at Yale University.

Table C.11: GEx Space Distances. GEx space inter-cluster distances in the WT organoid and fibroblast co-culture. Cells are coloured according to their relative distance values.

	ER Stress	E. Entero.	L. Entero.	Secret.	proCSC	CSC	TA2	revCSC	TA1	Fibroblast
ER Stress	0.000000	0.134136	0.087232	0.144213	0.161209	0.154001	0.553933	0.129671	0.187844	0.759904
E. Entero.	0.134136	0.000000	0.110047	0.166665	0.185258	0.177877	0.580909	0.153192	0.211384	0.797288
L. Entero.	0.087232	0.110047	0.000000	0.116684	0.139993	0.132073	0.543032	0.106004	0.164687	0.790359
Secret.	0.144213	0.166665	0.116684	0.000000	0.197554	0.189356	0.604088	0.162743	0.221462	0.864652
proCSC	0.161209	0.185258	0.139993	0.197554	0.000000	0.203334	0.596904	0.180319	0.237957	0.776085
CSC	0.154001	0.177877	0.132073	0.189356	0.203334	0.000000	0.592524	0.173165	0.230962	0.781164
TA2	0.553933	0.580909	0.543032	0.604088	0.596904	0.592524	0.000000	0.573983	0.630285	1.000000
revCSC	0.129671	0.153192	0.106004	0.162743	0.180319	0.173165	0.573983	0.000000	0.206725	0.787210
TA1	0.187844	0.211384	0.164687	0.221462	0.237957	0.230962	0.630285	0.206725	0.000000	0.835137
Fibroblast	0.759904	0.797288	0.790359	0.864652	0.776085	0.781164	1.000000	0.787210	0.835137	0.000000

Table C.12: LRT-KG Projection Space Distances. Projected LRT-KG inter-cluster distances in the WT organoid and fibroblast co-culture. Cells are coloured according to their relative distance values.

	ER Stress	E. Entero.	L. Entero.	Secret.	proCSC	CSC	TA2	revCSC	TA1	Fibroblast
ER Stress	0.000000	0.215563	0.145190	0.139818	0.201897	0.218083	0.632105	0.198838	0.277580	0.803968
E. Entero.	0.215563	0.000000	0.153977	0.147012	0.211975	0.227420	0.647468	0.209010	0.287032	0.828309
L. Entero.	0.145190	0.153977	0.000000	0.074063	0.143724	0.159109	0.587539	0.138559	0.217404	0.802999
Secret.	0.139818	0.147012	0.074063	0.000000	0.138364	0.153222	0.583433	0.132827	0.211676	0.803792
proCSC	0.201897	0.211975	0.143724	0.138364	0.000000	0.210823	0.621495	0.194830	0.271929	0.761176
CSC	0.218083	0.227420	0.159109	0.153222	0.210823	0.000000	0.640803	0.211221	0.288502	0.790591
TA2	0.632105	0.647468	0.587539	0.583433	0.621495	0.640803	0.000000	0.627242	0.703030	1.000000
revCSC	0.198838	0.209010	0.138559	0.132827	0.194830	0.211221	0.627242	0.000000	0.270369	0.807421
TA1	0.277580	0.287032	0.217404	0.211676	0.271929	0.288502	0.703030	0.270369	0.000000	0.873586
Fibroblast	0.803968	0.828309	0.802999	0.803792	0.761176	0.790591	1.000000	0.807421	0.873586	0.000000

Appendix D

Qin & Cardoso Rodriguez *et al.*, 2023

A Single-cell Perturbation Landscape of Colonic Stem Cell Polarisation

Xiao Qin [1*], Ferran Cardoso Rodriguez [1*], Jahangir Sufi [1], Petra Vlckova [1], Jeroen Claus [2], and Christopher J. Tape [1,✉]

[1]Cell Communication Lab, Department of Oncology, University College London Cancer Institute, 72 Huntley Street, London, WC1E 6DD, UK.; [2]Phospho Biomedical Animation, The Greenhouse Studio 6, London, N17 9QU, UK.; [*]These authors contributed equally to this work.

Cancer cells are regulated by oncogenic mutations and microenvironmental signals, yet these processes are often studied separately. To functionally map how cell-intrinsic and cell-extrinsic cues co-regulate cell-fate in colorectal cancer (CRC), we performed a systematic single-cell analysis of 1,071 colonic organoid cultures regulated by 1) CRC oncogenic mutations, 2) microenvironmental fibroblasts and macrophages, 3) stromal ligands, and 4) signalling inhibitors. Multiplexed single-cell analysis revealed a stepwise epithelial differentiation landscape dictated by combinations of oncogenes and stromal ligands, spanning from fibroblast-induced Clusterin (CLU)$^+$ revival colonic stem cells (revCSC) to oncogene-driven LRIG1$^+$ hyper-proliferative CSC (proCSC). The transition from revCSC to proCSC is regulated by decreasing WNT3A and TGF-β-driven YAP signalling and increasing KRAS^{G12D} or stromal EGF/Epiregulin-activated MAPK/PI3K flux. We find APC-loss and KRAS^{G12D} collaboratively limit access to revCSC and disrupt stromal-epithelial communication – trapping epithelia in the proCSC fate. These results reveal that oncogenic mutations dominate homeostatic differentiation by obstructing cell-extrinsic regulation of cell-fate plasticity.

Correspondence: *c.tape@ucl.ac.uk*

Highlights

- 1,071-condition single-cell transition map of colonic stem cell polarisation regulated by oncogenic and mircoenvironmental cues.
- Fibroblasts polarise WT colonic epithelia towards *Clu*$^+$ revCSC via TGF-β1 and YAP signalling.
- APC-loss and KRAS^{G12D} drive a *Birc5*$^+$, *Lrig1*$^+$, and *Ephb2*$^+$ proCSC fate via MAPK and PI3K.
- Oncogenic mutations disrupt stromal regulation of epithelial plasticity, trapping cells in the proCSC fate.

Introduction

The intestinal epithelium comprises multiple cell-types fulfilling the functions of nutrient absorption, waste elimination, and barrier protection [1]. In the healthy colon, a subpopulation of epithelial cells are maintained in a multipotent stem cell state by the pericryptal mesenchymal niche [2]. Stromal fibroblasts secrete paracrine ligands including WNT, EGF, Noggin, and R-Spondin-1 to maintain epithelial stemness and guide differentiation towards secretory and absorptive cells along the crypt [3]. In colorectal cancer (CRC), oncogenic mutations targeting *Apc*, *Kras*, *Braf*, *Smad4*, and/or *Trp53* cell-autonomously induce a crypt-progenitor phenotype in CRC cells [4]. Thus, in both the healthy colon and CRC, a subpopulation of epithelial cells are maintained in a stem-like state – albeit by different mechanisms.

Colonic epithelial stem cells are traditionally described as LGR5$^+$ OLFM4$^+$ crypt base progenitors [5]. However, recent single-cell studies of intestinal epithelia have identified additional multipotent cell-types, most notably Clusterin (CLU)$^+$ 'revival' or 'foetal' stem cells [6]. Revival stem cells can be induced following tissue damage to repopulate all epithelial cell-types but are otherwise rare in the homeostatic intestine [7]. Revival-like stem cells have also been implicated in CRC initiation [8], can be observed in developed CRC tumours in a patient-specific manner [9], and are emerging as putative drug-tolerant persister cells in CRC [10]. However, how combinations of oncogenic signals and microenvironmental cues regulate the polarisation of epithelia towards traditional or revival stem cells is unclear.

The CRC tumour microenvironment (TME) is a heterocellular system where cell-intrinsic oncogenic mutations and cell-extrinsic stromal and immunological signalling cues co-regulate epithelial cancer cells [11]. Stromal ligands and oncogenic mutations can activate common intracellular signalling pathways in colonic epithelia [12]. Canonically, both stromal WNT/R-Spondin-1 ligands and APC-loss hyper-activate β-catenin signalling, whereas EGF and KRAS/BRAF mutations stimulate the MAPK pathway [1]. As a consequence of their overlapping signalling mechanisms, oncogenic mutations must compete with stromal ligands during oncogenesis – yet how cell-intrinsic and cell-extrinsic cues co-regulate epithelial cell-fate remains elusive.

Here we describe a functional single-cell study exploring how cell-extrinsic and cell-intrinsic cues co-regulate colonic epithelial fate. Parallel perturbation analysis of >1,000 heterocellular organoid cultures using single-cell RNA-sequencing (scRNA-seq) and highly-multiplexed thiol-reactive organoid barcoding *in situ* (TOB*is*) mass cytometry (MC) [13] revealed that fibroblasts and oncogenic mutations induce distinct epithelial stem cell-fates

in colonic epithelia. We find that fibroblasts polarise epithelia towards slow-cycling CLU^{+} revival stem cells via TGF-β1 and YAP, whereas APC-loss, KRAS^{G12D}, and/or exogenous Epiregulin (EREG) shift cells towards a LRIG1^{+} hyper-proliferative fate that is dependent on PI3K signalling. APC-loss and KRAS^{G12D} collaboratively block cell-extrinsic regulation of epithelial plasticity by interrupting stromal-epithelial communication, trapping CRC cells in a cancerous state. Despite the dominance of oncogenes over epithelial plasticity, we find that CRC organoids can still access revival stem cells, but this requires high cell-extrinsic activation of YAP via TGF-β1 in parallel with reduced PI3K signalling.

These results demonstrate that colonic epithelia exist on a continuous differentiation landscape where oncogenic mutations and stromal cues compete for epithelial identity – but oncogenes eventually dominate by blocking the stromal regulation of cell-fate plasticity.

Results

Oncogenic and Stromal Cues Differentially Regulate Colonic Epithelia

To directly compare how CRC oncogenic mutations and stromal cells regulate colonic epithelial differentiation, we performed a multivariate scRNA-seq analysis of wild-type (WT), *shApc* (A), *shApc* and *Kras$^{G12D/+}$* (AK), and *shApc*, *Kras$^{G12D/+}$* and *Trp53$^{R172H/-}$* (AKP) colonic organoids, in monoculture or co-cultured with colonic fibroblasts and/or macrophages (Figure 1A). Fibroblasts are established regulators of intestinal epithelia [14] and macrophages are the most profuse leukocytes in the colon [15]. WT epithelia cultured with exogenous WNT3A, EGF, Noggin, and R-Spondin-1 (WENR) (commonly used to grow colonic organoids) were included as a defined mesenchymal niche factor control.

Following scRNA-seq, epithelial cells, fibroblasts, and macrophages were resolved by Leiden clustering [16], visualised by PHATE (Potential of Heat-diffusion for Affinity-based Trajectory Embedding) [17] (Figure 1B), and cell-type-specific transcriptional changes were summarised by principal component analysis (PCA) (Figure 1C). Epithelial transcriptomes are differentially regulated by both CRC mutations (PC1, 26%) and microenvironmental cues (PC2, 22%), with A, AK, and AKP mutations progressively dysregulating their transcriptomic profiles. However, we found fibroblasts can only regulate WT and A epithelial cells (Figure 1C). Although WENR ligands are thought to mimic a healthy stromal niche [18], WT organoids + WENR ligands transcriptionally align with AK mutant organoids (not WT+fibroblasts as might be expected), indicating this widely used colonic organoid culture media induces a partial CRC-like transcriptome in WT epithelia (Figure 1C).

Colonic fibroblasts clustered into CD34^{hi} and CD34^{lo} subpopulations mimicking *in vivo* stromal heterogeneity [19, 20] (Figure S1A). CD34^{hi} and CD34^{lo} fibroblasts did not differentially regulate colonic epithelia (Figure S1B) and were subsequently treated as a heterogenous mesenchymal population. We found fibroblast and macrophage transcriptomes were only regulated by co-culture with heterotypic cells but not altered by epithelial genotypes (Figures 1C, S1C-D).

Oncogenic Mutations and Fibroblasts Polarise Epithelia Towards Distinct Stem Cell-Fates

Epithelial cells from all conditions were integrated by reciprocal PCA (RPCA) [16], projected onto a shared PHATE embedding, and clustered into multiple cell-fates, including stem populations, transit amplifying (TA) cells, cells under ER stress, goblet and deep crypt secretory (DCS) cells, and early or late enterocytes (Figure 1D). Stem clusters contain high signalling entropy (indicative of pluripotency) [21] and act as origins for RNA velocity streams [22] that transition towards differentiated cells (Figures 1E, S2E).

Differential abundance testing [23] of co-culture and CRC monoculture conditions against WT monocultures revealed that fibroblasts, macrophages, and CRC mutations have markedly different effects on epithelial cell-fate determination (Figure 1F-H). Fibroblasts enrich a distinct stem cell population characterised by high expression of epithelial progenitor genes *Clu*, *Sox9*, *Cd44*, and *Cldn4* (Figures 1I). These fibroblast-induced stem cells are transcriptionally similar to 'foetal' [24, 25] or 'revival' stem cells (revSCs) [7] of the small intestine (S2A) and are hereafter referred to as 'revival colonic stem cells' (revCSC).

In contrast, A, AK, and AKP mutations progressively polarise epithelia towards a hyper-proliferative colonic stem cell-fate, hereafter named proCSC (Figure 1G, H). proCSCs express *EphB2*, *Birc5* (*Survivin*), *Lrig1*, *Hmgb2*, and *Rrm2* and are highly mitotic (*Stmn1^{+}*, *Mki67^{+}*, and *Ccnb1^{+}*) (Figure 1I). In addition, proCSCs are transcriptionally comparable to stem cells observed in mouse and human CRC (Figure S2A). Both revCSC and proCSC are present in WT organoids at low levels alongside traditional *Lgr5^{+}* colonic stem cells, hereafter named CSC (Figure S2B). We found CSC are also enriched by A, AK, and AKP genotypes, but to a lesser extent than proCSC, and CSC gene signatures are less common in CRC (Figure S2A).

We found that fibroblasts can only induce revCSC in WT and *shApc* epithelia, but not when cells contain both *shApc* and *Kras$^{G12D/+}$* (Figure 1H). Conversely, proCSCs are enriched in all A, AK, and AKP organoids irrespective of fibroblasts or macrophages, suggesting oncogenic mutations are dominant over microenvironmental signalling. WENR ligands hyper-polarise WT epithelia towards all stem and TA cell-types, with very few cells retaining secretory or absorptive identities (Figures 1H-I, S2B). WT epithelia also show higher RNA velocity vector lengths relative to CRC cells (Figure S2C-D), suggesting that

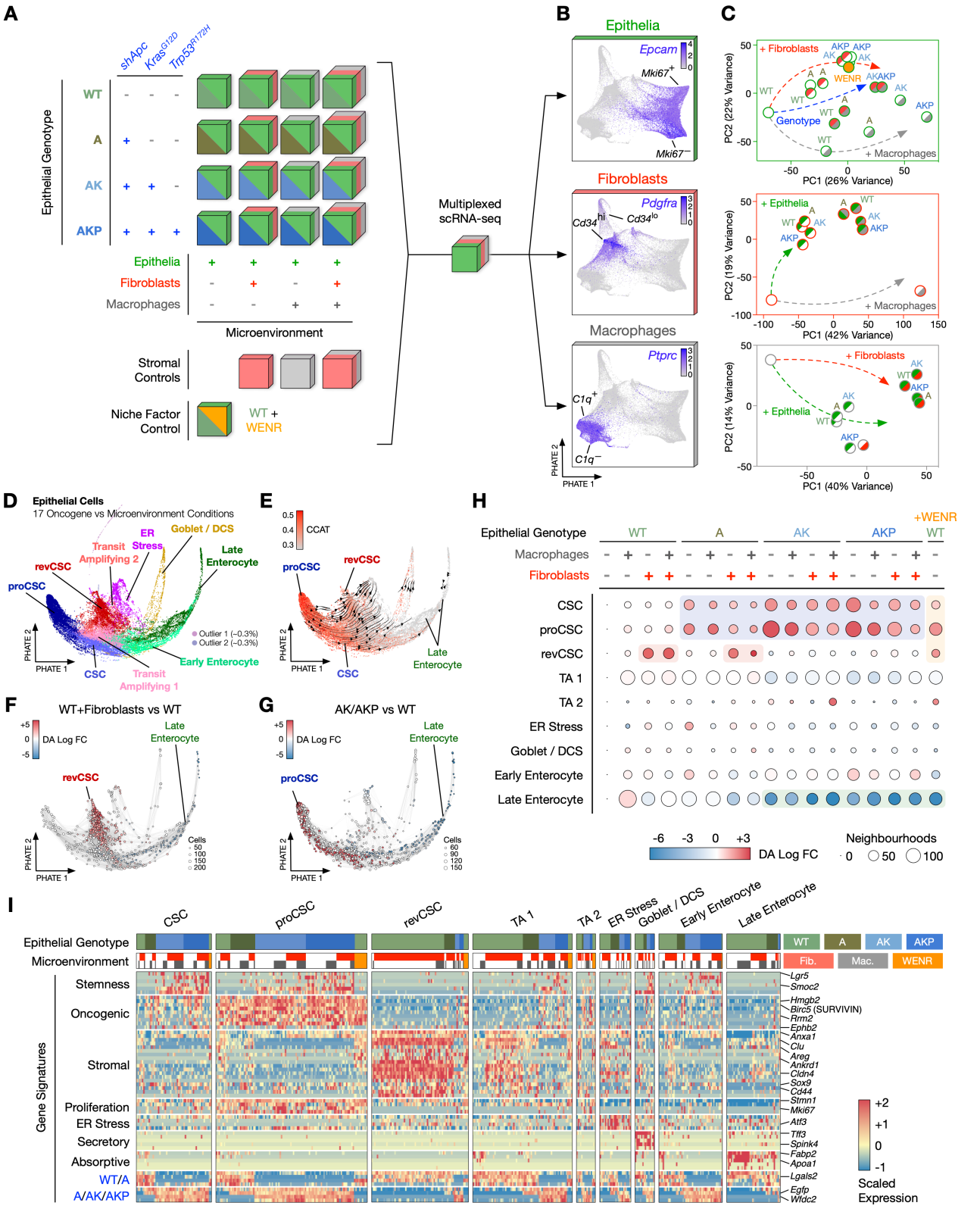


Figure 1. Oncogenes and Fibroblasts Differentially Regulate Colonic Epithelia. A) Multivariate scRNA-seq experimental design. WENR ligands were removed from all experimental conditions except for the niche factor control to ensure cell-cell signalling was not dominated by exogenous recombinant proteins (see Methods). **B)** Single-cell PHATE embedding illustrating epithelial cells, fibroblasts, and macrophages. **C)** PCAs of epithelial, fibroblast, and macrophage transcriptomes regulated by by organoid genotype and microenvironment. **D)** PHATE embedding of 29,452 epithelial cells from the 17 organoid conditions coloured by cell-type clusters. **E)** Epithelial PHATE coloured by CCAT score and overlaid with velocity streams (arrows). **F)** Epithelial PHATE overlaid with differentially abundant (DA) neighbourhoods in WT organoid + fibroblast co-cultures compared with WT organoid monocultures. **G)** Epithelial PHATE overlaid with DA neighbourhoods in AK/AKP organoid monocultures compared with WT organoid monocultures. **H)** Dot plot of epithelial clusters across organoid cultures coloured by log fold-change (Log FC) in neighbourhood abundance and sized by the number of neighbourhoods detected. **I)** Gene expression signatures of epithelial clusters. WENR, WNT3A, EGF, Noggin, and R-Spondin-1. CSC, colonic stem cell. proCSC, hyper-proliferative CSC. revCSC, revival CSC. DCS, deep crypt secretory cell. CCAT, correlation of connectome and transcriptome. TA, Transit amplifying cell.

oncogenic mutations reduce epithelial plasticity. While macrophages can alter epithelial gene expression (Figure 1C), macrophages do not regulate the abundance of epithelial cell-types (Figure 1H). In summary, multivariate scRNA-seq revealed that fibroblasts, CRC mutations, and WENR ligands polarise epithelia towards a de-differentiated progenitor state – with fibroblasts and oncogenes inducing distinct revCSC and proCSC fates.

WNT3A Polarises Epithelia to revCSC and Oncogenic Mutations to proCSC

Multivariate scRNA-seq demonstrated that cell-extrinsic ligands and cell-intrinsic mutations differentially regulate epithelial cell-fate, but could not describe how individual ligands and mutations co-regulate differentiation. To functionally explore epithelial polarisation, we performed a highly-multiplexed TOB*is* MC [12] combinatorial study focusing on the three axes hypothesised to regulate epithelial cell-fate: 1) microenvironment (+/- fibroblasts), 2) stroma-mimicking ligands (+/- WNT3A, +/- EGF, +/- Noggin, +/- R-Spondin-1), and 3) oncogenic mutations (+/- *shApc*, +/- *Kras*$^{G12D/+}$, +/- *Trp53*$^{R172H/-}$) (Figure 2A). Each organoid culture was performed in triplicate, barcoded *in situ* using 126-plex TOB*is* [13], pooled, dissociated into single cells, stained with a panel of 45 rare-earth metal-labelled antibodies (spanning epithelial differentiation markers identified by scRNA-seq, cell-state markers, and PTM signalling nodes [12]) (Table S1), and analysed by MC. Following debarcoding [26], QC, and cell-type-specific gating, we obtained 6 million cells from 390 organoid/fibroblast cultures (570 cell-type-specific single-cell datasets) (Figure 2B-D).

In agreement with scRNA-seq, analysis of 360 epithelial single-cell profiles confirmed that fibroblasts induce CLU^{+} revCSC in WT epithelia (Figures 2C,E, S3A), whereas oncogenic mutations induce hyper-proliferative LRIG1^{+}, EPHB2^{+}, and SURVIVIN^{+} proCSC while blocking access to revCSC (Figures 2B, E, S3A).

The effect of WENR ligands on epithelial differentiation is highly dependent on genotype (Figure 2D, F). For example, when WT or A organoids are treated with R-Spondin-1 alone, no distinct shift in cell-signalling or cell-state is observed. However, when K (*Kras*$^{G12D/+}$) or KP (*Kras*$^{G12D/+}$, *Trp53*$^{R172H/-}$) cells are treated with R-Spondin-1, they undergo a dramatic S-phase entry and phenocopy AK and AKP genotypes (Figures 2F, S3B). This suggests that KRAS^{G12D} fundamentally rewires how epithelial cells respond to canonical β-catenin signalling (via stromal R-Spondin-1 or APC-loss) to bias epithelia towards proCSC. By contrast, WNT3A upregulates CLU in WT, A, K, and KP epithelia but only shows a very minor effect on cells containing both *shApc* and *Kras*$^{G12D/+}$ (Figures 2F, S3C), indicating that APC-loss and oncogenic KRAS dominate over epithelial response to exogenous cues, blocking access to revCSC and entrapping epithelia in the proCSC fate.

Despite their origin as stroma-mimicking cues, we found that WENR ligands regulate epithelia very differently from fibroblasts (Figure S3D-F). Purified WNT3A enriches quiescent revCSCs with low mitogenic PTM signalling activity. Conversely, fibroblasts induce SOX9^{+}, pRB [S807/S811]$^{+}$ revCSCs with high levels of MAPK (pERK1/2 [T202/Y204], pMKK3/6 [S189/S207], pMAPKAPK2 [T334], and pP90RSK [T359]) and TGF-β (pSMAD2/3 [S465/S467]) signalling (Figure S3F). This suggests that fibroblast-induced revCSCs are distinct from those regulated by WNT3A alone and that the communication between stromal and epithelial cells is more diverse than just WENR ligands.

Oncogenic Mutations and Stromal Ligands Regulate Epithelia Across a Continuous Differentiation Trajectory

To understand how organoid monocultures are regulated by WENR ligands, we analysed WT, A, K, AK, KP, and AKP organoids treated +/- WNT3A, EGF, Noggin, and R-Spondin-1 (180 single-cell profiles). This analysis revealed that colonic epithelial differentiation exists on a multivariate continuum where stromal ligands and oncogenic cues compete for epithelial fate (Figure 3A). We observed a clear fate-transition trajectory of epithelial differentiation dictated by oncogenes and ligands, spanning from WNT3A-driven WT revCSC, through an equilibrium of balanced stem cell identities and enterocyte differentiation, to oncogene-dominant proCSC (Figures 3A, S4A-B). Crucially, WNT3A can drive epithelia towards the revCSC fate when only one oncogenic-driver is present, but the combination of APC-loss and KRAS^{G12D} traps epithelia in the proCSC state that is largely unresponsive to all WENR ligands (Figure 3A).

We found that the regulation of revCSC by WNT3A is also heavily influenced by parallel EGF signalling. For example, WNT3A alone leads to quiescent CLU^{+} revCSC in WT epithelia, but if WNT3A is combined with EGF, cells maintain cell-cycle activity and achieve an equilibrium of stem identities (Figures 2F, S4B). The transition between revCSC and equilibrium can be clearly observed across a WNT3A vs EGF gradient and fine-tuned by altering the ratio between EGF and WNT3A concentrations (Figure S4C-F). This suggests that the access to revCSC is controlled by competing signalling flux downstream of WNT3A and EGF.

Consistent with the hypothesis that revCSC and proCSC are regulated by different signalling pathways, TOB*is* MC demonstrated that revCSCs have low cell-cycle activity and high pGSK-3β [S9], whereas epithelia in the equilibrium state display activated pNDRG1 [T346] and pMKK3/6 [S189/S207]. In contrast, proCSC lose cytokeratin expression and have very high levels of PI3K signalling (e.g. pAKT [T308], pPKCα [T497], and p4E-BP1 [T37/T46]) (Figure 3B). The continuous regulation of epithelia by CRC mutations and ligands can be orthogonally depicted in a genotype-anchored scaffold map [27], where revCSC-enriched WT+WNT3A transition into

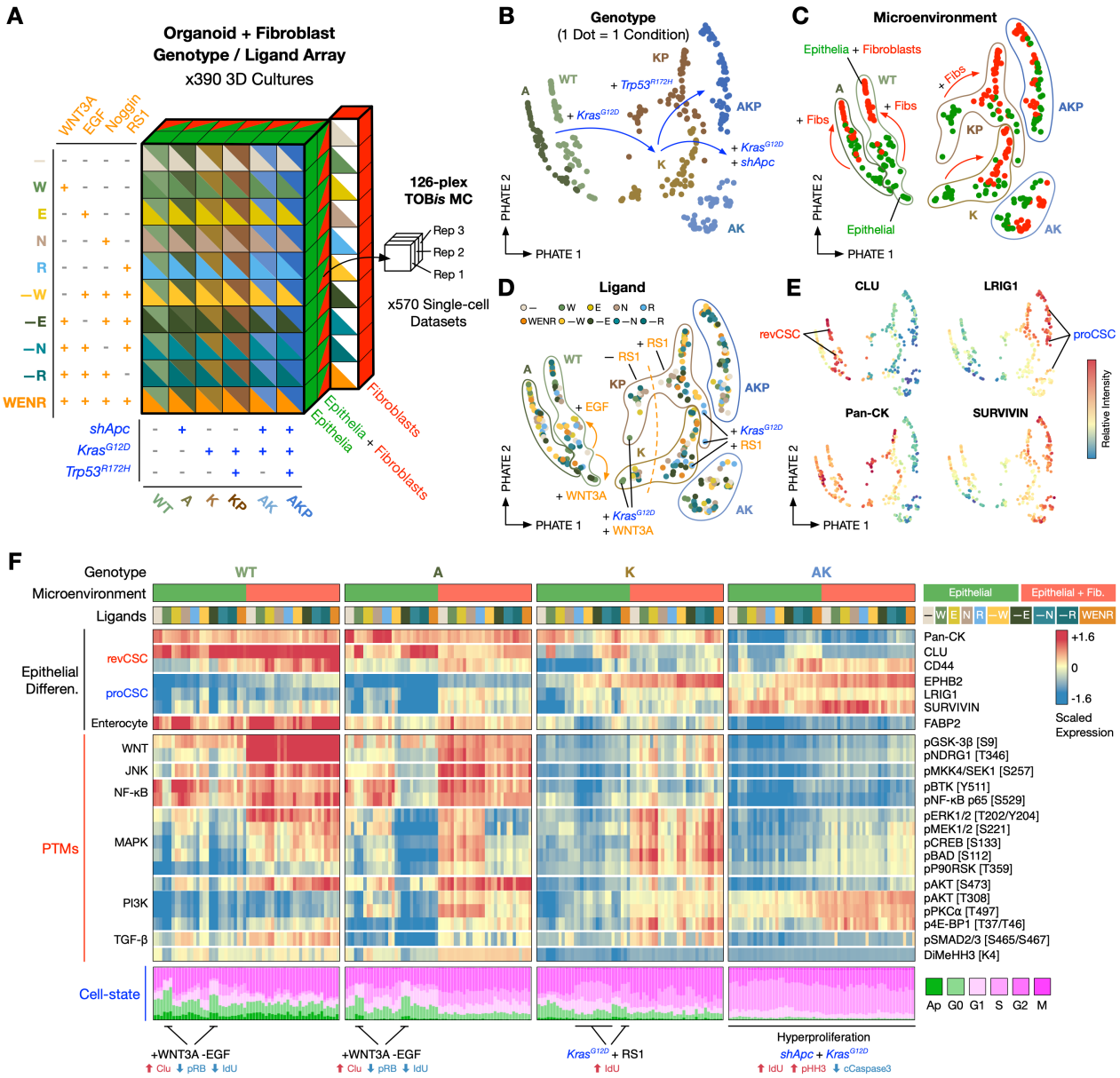


Figure 2. Regulation of Colonic Epithelia by Epithelial Genotypes, Fibroblasts, and WENR Ligands. A) TOB*is* MC multidimensional array comprising epithelial genotypes, fibroblasts, and WENR ligands (570 single-cell datasets). **B-E)** EMD-PHATE of 360 organoid cultures coloured by genotype, microenvironment, WENR ligands, and EMD scores of epithelial cell markers. One dot = one condition. **F)** Relative expression of epithelial markers, PTMs, and cell-state markers regulated by genotypes, fibroblasts, and/or WENR ligands. One column = one condition. MC, mass cytometry. Fib., fibroblasts. Ap., Apoptotic.

proCSC-dominant K+R-Spondin-1 and AK conditions in a stepwise manner (Figure 3C). The regulation of epithelial stem cell-fate by WENR ligands can be described by simple genotype-specific Boolean logic models (Figure 3D). These models reveal that while WT epithelia are highly sensitive to cell-extrinsic reprogramming, *shApc* and $Kras^{G12D/+}$ progressively limit epithelial plasticity and cell-intrinsically trap epithelia in the proCSC fate.

Oncogenic Mutations Inhibit Fibroblast-Epithelia Signalling

As epithelial differentiation cannot be regulated by fibroblasts in the context of *shApc* and $Kras^{G12D/+}$ (Figures 1H, 2C), we hypothesised oncogenic mutations might disrupt stromal-epithelial signalling. To test this, we performed ligand-receptor cell-cell communication analysis [28] of WT, A, AK, and AKP organoid+fibroblast co-culture scRNA-seq datasets.

Given their established role in microenvironmental cell-cell communication, fibroblasts unsurprisingly demonstrate high 'outgoing' signalling (i.e., express numerous ligands and extracellular matrix (ECM) components). By contrast, WT epithelia display a dominant 'incoming' signalling potential (i.e., express many receptors) (Figure 4A). This dichotomy suggests that heterocellular signalling in the healthy colon is largely unidirectional from fibroblasts to epithelial cells. We found that revCSC and the closely affiliated TA 1 and TA 2 clusters are responsible for much of the 'incoming' signalling potential of WT epithelia, indicating these cell-types are hyper-

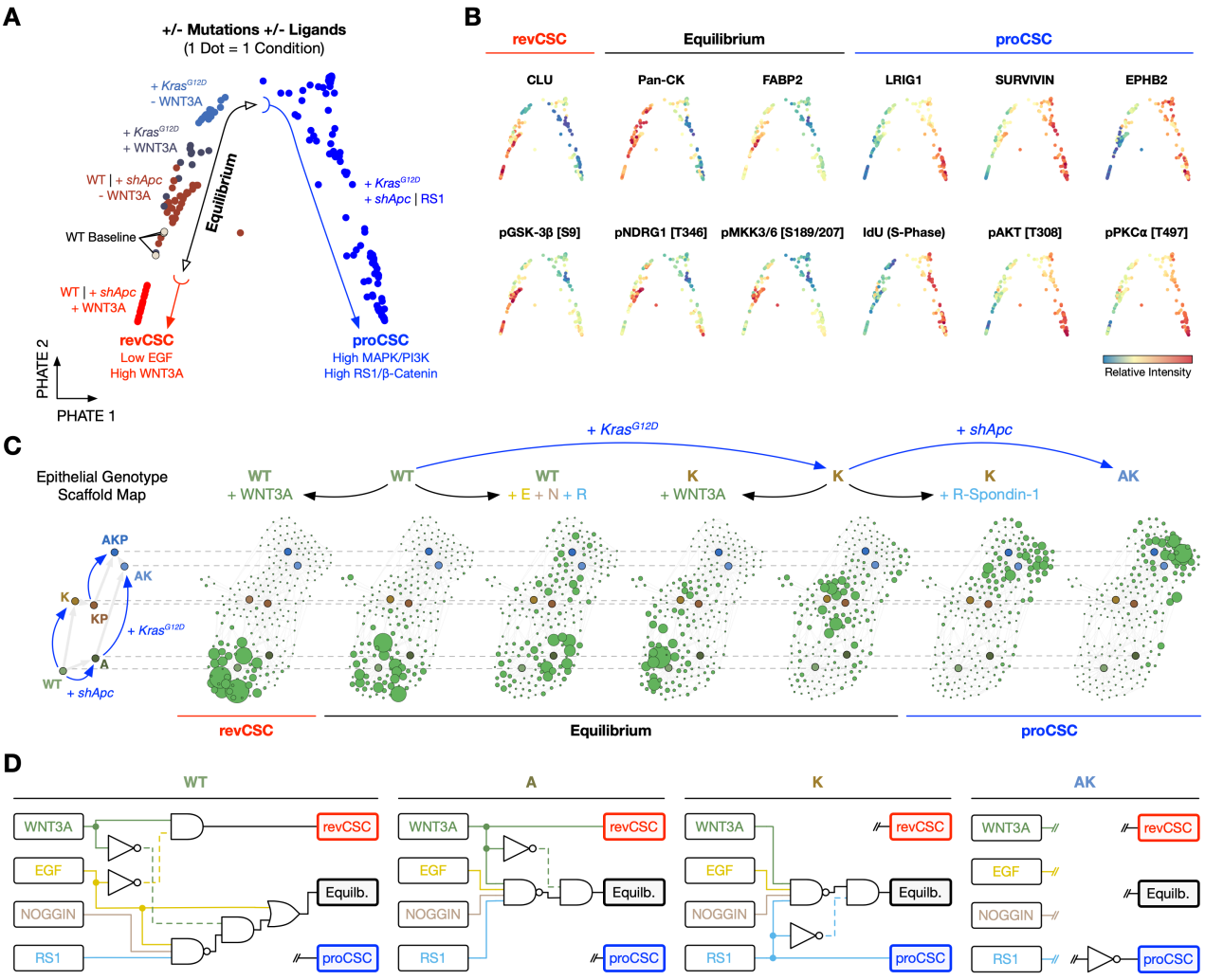


Figure 3. Stepwise Transition from revCSC to proCSC Regulated by Oncogenic Mutations and Ligands. A) EMD-PHATE of 180 organoid single-cell datasets regulated by epithelial genotype and WENR ligands. One dot = one condition. **B)** EMD-PHATE coloured by EMD scores of revCSC, equilibrium, and proCSC markers. **C)** Epithelial genotype scaffold maps with organoid monoculture landmarks and genotype+ligand overlays. **D)** Boolean logic models for genotype-specific regulation of colonic stem cells (CSC) by WENR ligands. revCSC, revival CSC. Equilb., equilibrium. proCSC, hyper-proliferative CSC. RS1, R-Spondin-1.

sensitive to cell-extrinsic regulation by fibroblasts. In contrast, proCSC are the least receptive of all epithelial cells, suggesting proCSC are more reliant on cell-intrinsic signalling (Figure 4A).

Cell-cell communication analysis revealed that fibroblasts form putative paracrine and juxtacrine interactions with WT and A cells, which are often lost in AK and AKP genotypes (Figure 4B). For example, WT and A organoids show intact NRG1, EREG, IGF, and TGF-β signalling with fibroblasts, but these cell-cell interactions are undetectable in AK and AKP cells, due to the down-regulation of epithelial signal receptors (Figures 4B-C, S5A-C).

Ligand-receptor analysis is increasingly used to generate putative cell-cell communication models in heterocellular systems [29], yet these computational hypotheses are rarely experimentally validated. To functionally test how oncogenic mutations regulate stromal-epithelial communication, we performed a systematic TOB*is* MC study of epithelial differentiation, cell-state, and PTM signalling in WT, A, K, KP, AK, and AKP organoids treated with stromal ligands identified by ligand-receptor analysis as WT homeostatic regulators (WNT5A, SEMA3A, TGF-β1, TGF-β2, IGF, NRG1, EREG, and OPN (*Spp1*)) (Figure 4B-C).

Single-cell MC analysis of 204 organoid cultures revealed that WT and A epithelia can be polarised towards revCSC by WNT3A or TGF-β1, whereas ERBB signalling via EGF, EREG, or NRG1 pushed cells towards the proCSC fate. This suggests that stem cell polarisation can be recapitulated by fibroblast-secreted ligands independent of stromal-epithelial contact or fibroblast-driven ECM remodelling. In contrast, ligands fail to regulate epithelia containing both *shApc* and *Kras*$^{G12D/+}$ (Figures 4D, S5D-F). The resistance to external signalling cues of AK/AKP epithelia mimics the diminishing stromal-epithelial communication predicted by ligand-receptor analysis (Figure 4A-C) and is reminiscent of their unresponsiveness to WENR ligands (Figure 3D). Collectively, this analysis suggests that the combination of APC-loss and oncogenic KRAS^{G12D} decouple epithelial cells from

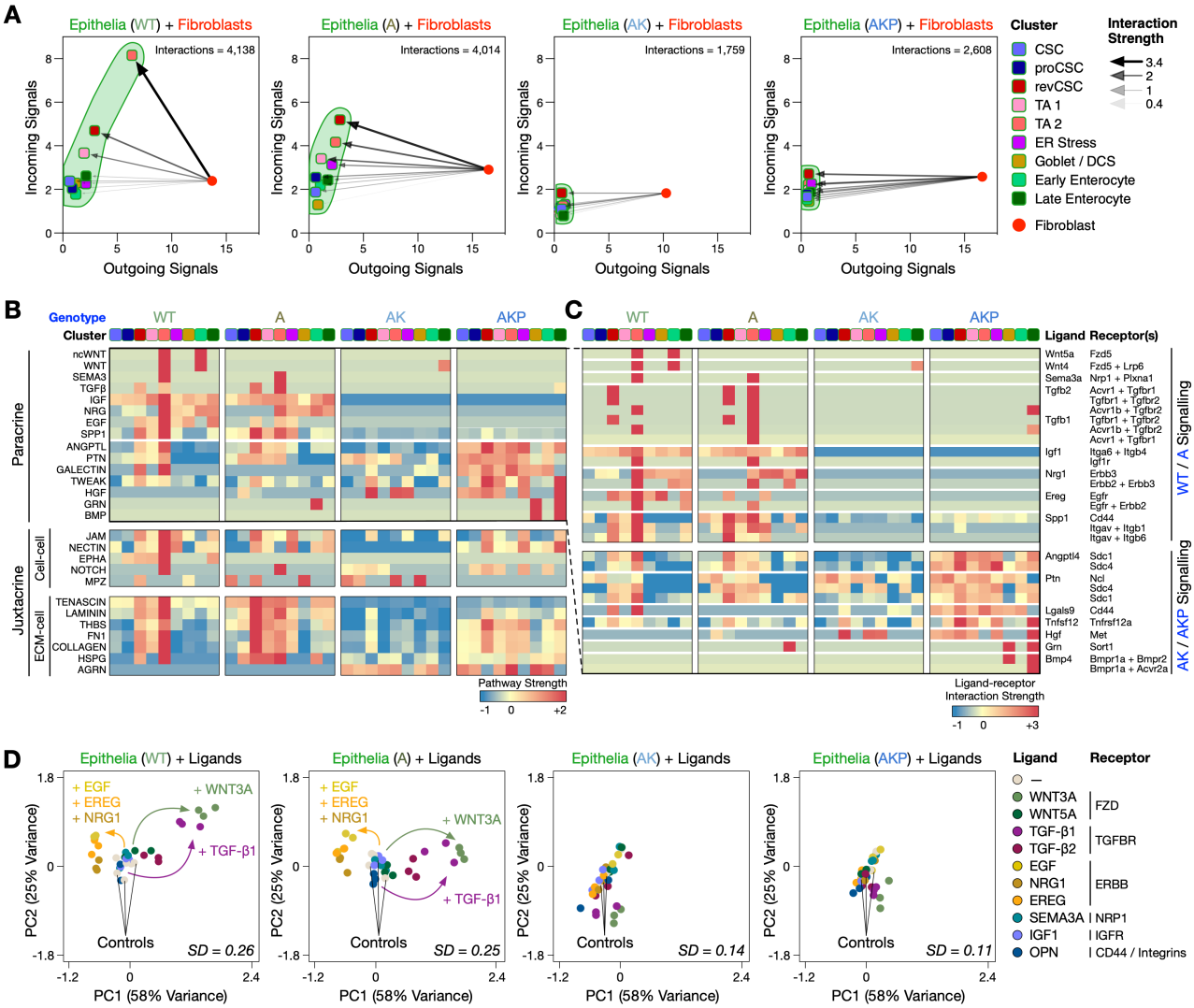


Figure 4. Oncogenic Mutations Disrupt Stromal-epithelial Communication. A) Outgoing and incoming communication probability (interaction strength) from fibroblasts to epithelia across organoid genotypes. **B-C)** Predicted paracrine and juxtacrine communication summarised at the pathway and ligand-receptor interaction level. **D)** EMD-PCA of epithelial regulation by exogenous ligands across the genotypes (138 single-cell datasets). One dot = one condition. SD, standard deviation of the distribution of EMD scores for each genotype.

homeostatic intercellular signalling – with CRC cells becoming 'bad listeners' in the tissue microenvironment.

revCSC and proCSC are Regulated by Competing Signalling Pathways

As epithelial cells are co-regulated by cell-intrinsic and cell-extrinsic cues across an integrated differentiation trajectory, we hypothesised different signalling pathways might compete to control epithelial cell-fate. To determine the signalling hubs regulating revCSC and proCSC polarisation, we performed an extensive single-cell cue-signal-response perturbation assay spanning: 1) CRC oncogenic mutations (*shApc* and *Kras*$^{G12D/+}$), 2) stem cell polarisation ligands (WNT3A, EREG, and TGF-β1), and 3) inhibitors targeting: β-catenin (ICG-001), GSK-3β (CHIR99021), MEK (Trametinib), PI3K (GDC-0941), FAK (PF-573228), SRC (Dasatinib), YAP (CA3), and SMAD3 (SIS3) (Figure 5A).

Analysis of 432 single-cell MC organoid profiles confirmed that WT, A, and K epithelia can be polarised towards revCSC by WNT3A or TGF-β1 and to proCSC by EREG. However, organoids containing both *shApc* and *Kras*$^{G12D/+}$ showed limited response to ligands and largely retained their proCSC identify (Figures 5B-E, S6A-B). While the ligand-effect is genotype-specific, signalling inhibitors can disrupt the polarisation of proCSC and revCSC across all genotypes, with several interesting examples of ligands and inhibitors collaborating to regulate epithelial cell-fates (Figures 5D, S6D-K).

To rank the polarisation of proCSC and revCSC by genotypes, ligands, and inhibitors across a shared regulation landscape, we established a relative stemness (RS) score by calculating the single-cell expression ratio between LRIG1 and CLU for each organoid culture (Figure 5F). In this space, WT+TGF-β1 have a low RS score, indicating enrichment of CLU^{+} revCSCs, whereas AK have a high RS score and are dominated by LRIG1^{+} proCSCs (Figures 5F, S6C). The differential polarisation

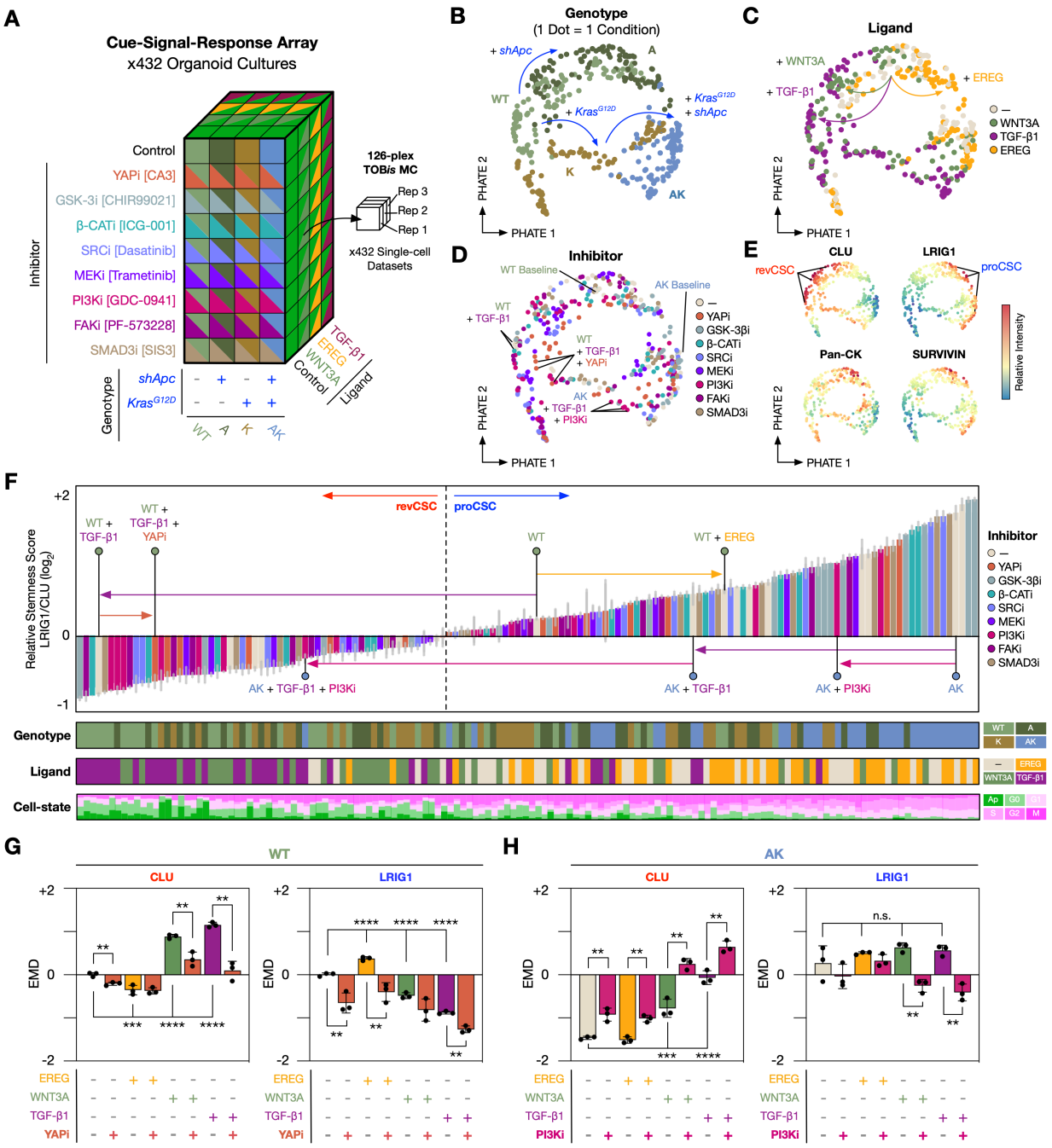


Figure 5. revCSC and proCSC are Regulated by Competing Signalling Nodes. A) Cue-signal-response organoid array experimental design. **B-E)** EMD-PHATE of 432 organoid cultures coloured by genotype, ligand, inhibitor, and EMD scores of epithelial cell markers. One dot = one condition. **F)** Ranked relative stemness score (log_2-transformed single-cell expression ratio between LRIG1 and CLU) across all conditions in the cue-signal-response array annotated by epithelial genotypes, exogenous ligands, and organoid cell-state. Error bars = SD. **G)** EMD scores for CLU and LRIG1 across WT organoid culture conditions. **H)** EMD scores for CLU and LRIG1 across AK organoid culture conditions. (**, $p < 0.01$; ***, $p < 0.001$; ****, $p < 0.0001$. n.s., not significant. Ordinary one-way ANOVA with Holm-Šídák's multiple comparisons test between untreated and ligand controls. Two-tailed unpaired *t*-test for inhibitor treatments). Error bars represent SD.

of proCSC and revCSC can therefore be captured using shifts in the RS score (Figure 5F).

Surprisingly, SMAD inhibition did not alter TGF-β1 regulation of revCSC (Figure S6F). However, both TGF-β1 and WNT3A regulation of revCSC could be partially reversed by YAP inhibition, suggesting revCSC is a YAP-dependent cell-fate (Figures S2A, 5F-G, and S6G). In contrast, although treating AK organoids with either TGF-β1 or PI3Ki alone caused a decrease in RS score (with the epithelial population still dominated by proCSCs), we found treatment of AK organoids with PI3Ki and TGF-β1 enabled epithelia to enter a revCSC-dominant state. This suggests that CRC cells can access revCSC, but the transition requires high TGF-β1 and low PI3K signalling (Figures 5F, 5H, and S6I). Collectively, the single-cell cue-signal-response perturbation array revealed that colonic stem cell plasticity is generally resilient, but cells can transition between revCSC and proCSC by re-balancing competing signalling flux in YAP, PI3K and MAPK path-

ways, even in CRC organoids. We found that YAP is a central regulator of revCSC, while PI3K and MAPK are important for maintaining the proCSC identity.

Single-cell Landscape of Colonic Epithelial Cell-fate Plasticity

In 1957, C.H. Waddington published his famous illustration of cellular differentiation, depicting pluripotent cells rolling down a landscape into valleys of terminal differentiation [30]. While an evocative metaphor in developmental biology, this conceptual model has not been clearly demonstrated with real data. However, recent computational advances in global-structure embeddings [17], differentiation potency metrics [21], and local differentiation-rate predictions [22] now provide the component elements to reconstruct Waddington-like embeddings from scRNA-seq data.

To visualise single-cell colonic epithelial differentiation on a Waddington-like landscape, we combined the global cellular relationships captured by PHATE [17] as 'longitude and latitude' axes, with an integrated Valley-Ridge (VR) score to represent pluripotent 'altitude'. The VR score is defined as the sum of two components per cluster: CCAT signalling-entropy [21] and RNA velocity [22]. At a cluster's centre, the VR score is solely determined by the median CCAT. However, the VR scores at the cluster periphery were augmented by weighting the inverse of RNA velocity component and the scaled distance from the cluster centre to model rates of local transcriptional change. This method reconstructs a data-driven estimate of Waddington-like landscapes where the altitude captures the differentiation potential of a cell population, with the valley-ridge topology delineating local plasticity (Figure 6A).

When WT colonic epithelia are projected onto this embedding, stem cells occupy high positions in the landscape, with TA cells descending into a central valley before diverging into terminally differentiated secretory and absorptive cells. When WT epithelia communicate with fibroblasts, the TA valley erodes as cells access revCSC. In contrast, CRC mutations *shApc* and *Kras*$^{G12D/+}$ resculpt the entire landscape, trapping most cells in the proCSC fate by restricting their differentiation potential (Figure 6A).

The functional perturbation experiments described in this study support a signalling model that underpins each landscape (Figure 6B). In homeostatic WT epithelia, WNT3A, EREG, and R-Spondin-1 drive balanced β-catenin, MAPK, PI3K, and YAP signalling to enable an equilibrium of stem and terminally differentiated cell-fates. When exposed to fibroblast-derived TGF-β1, WT cells become dominated by the YAP signalling flux, have minimal MAPK and PI3K activity, and are therefore polarised towards revCSC. By contrast, APC-loss and KRAS^{G12D} hyper-activate cell-intrinsic β-catenin, MAPK, and PI3K signalling, while simultaneously downregulating receptor expression to decouple epithelia from cell-extrinsic regulation. This limits CRC access to revCSC and traps cells in the proCSC fate. CRC cells can only escape proCSC through high TGF-β1 and low PI3K – tipping the signalling balance back towards revCSC. These observations demonstrate that colonic epithelia exist on an integrated differentiation landscape that can be traversed by co-regulating core signalling hubs, either through cell-intrinsic mutations or cell-extrinsic ligands.

Discussion

Single-cell technologies can describe cell-type-specific regulation of differentiation and cell-cell communication [31, 32, 33]. In this study, we utilised both multiplexed scRNA-seq and high-throughput MC to functionally map how oncogenic mutations and stromal cues co-regulate colonic epithelia across a continuous polarisation landscape. By analysing >1,000 organoid cultures at single-cell resolution, we identify a stepwise cell-fate trajectory spanning from fibroblast-induced revCSC through an equilibrium of balanced differentiation to oncogene-driven proCSC. While scRNA-seq provides in-depth description of colonic epithelial differentiation and proCSC/revCSC polarisation, multiplexed TOB*is* MC allows comprehensive functional interrogation of cell-intrinsic and -extrinsic cues regulating each cell-fate.

The intestinal stroma comprises a heterogenous population of fibroblasts that regulate the intestinal stem cell niche [2]. In the colonic epithelium, CD34^{hi} fibroblasts located at the crypt bottom are a major source of WNT2B, GREM1, and R-Spondin-1, contributing to both homeostatic stem cell maintenance and tissue regeneration following injury [19]. In contrast, CD34^{lo} fibroblasts reside around upper crypts, show lower expression of WNT2B/GREM1 but higher expression of BMPs, thereby providing a permissive environment for epithelial differentiation [7, 20]. The fibroblasts used in this study contain both CD34^{hi} and CD34^{lo} cells – mimicking *in vivo* heterogeneity (Figure 1B). Both CD34^{hi} and CD34^{lo} fibroblast subpopulations showed comparable polarisation of revCSC (Figure S1B), suggesting the stromal-epithelial communication in organoid co-cultures may be dominated by TGF-β1 signalling (Figure 6B). While this study uses healthy colonic fibroblasts to model homeostatic signalling, it is possible cancer associated fibroblasts (CAFs) will communicate differently with epithelial cells, particularly in CRC. Future cell-cell communication studies between CAF sub-types [34] and defined epithelial genotypes could uncover exceptions to the signalling models described here and therefore provide novel avenues for therapeutic intervention in CRC.

WNT3A is considered a canonical WNT ligand that activates APC/β-catenin signalling, promotes cell proliferation, and reinforces stem cell identity in the intestinal epithelium [35]. It is therefore widely used in colonic organoid culture to compensate for the absence of Paneth cell-derived WNT3A compared with

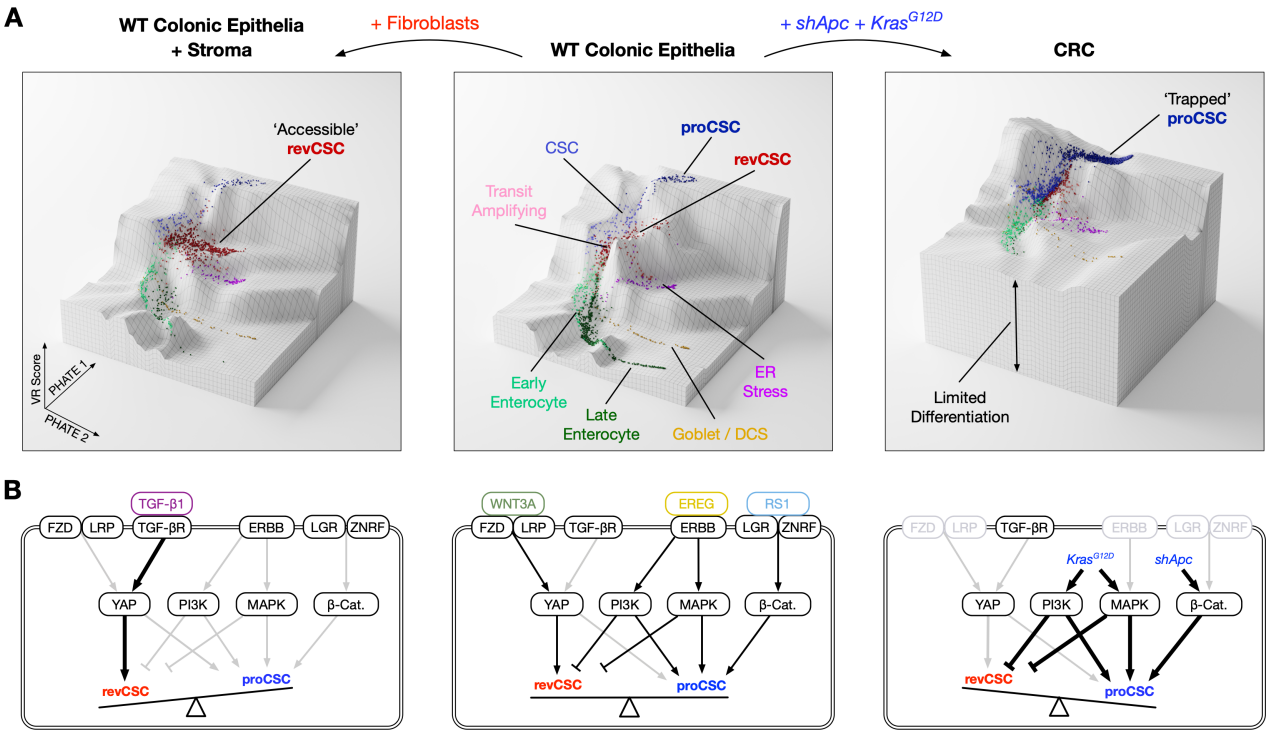


Figure 6. Fibroblast- and Oncogene-driven Waddington-like Single-cell Landscapes. A) Integrating PHATE and Valley-Ridge (VR) score enables Waddington-like embeddings of scRNA-seq data. Landscapes illustrate how WT epithelia differentiate from high signalling-entropy stem cells, through TA cells, into secretory and absorptive cells. Fibroblasts enable WT epithelia to access revCSC while retaining secretory and absorptive differentiation. In contrast, *shApc* and $Kras^{G12D/+}$ limit differentiation and trap cells in the proCSC state. **B)** Data-driven signalling models underpinning the transition from revCSC to proCSC. Arrow colour indicates pathway activation (black on, grey off), while arrow weight depicts relative signalling flux.

the small intestine [36]. Surprisingly, we found WNT3A alone polarised WT epithelia towards the slow-cycling revCSC fate. Moreover, *shApc* cannot induce revCSC cell-autonomously, indicating revCSC is not immediately downstream of canonical APC/β-catenin signalling (Figure S3A). Our data suggests that WNT3A drives the polarisation to revCSC via YAP (not β-catenin) (Figure 5G), and homeostatic differentiation requires balanced EGF and WNT3A signalling (Figure S4E-F). WT organoids cultured with WENR ligands are enriched for both proCSC and revCSC while depleted of secretory cells and enterocytes (Figures 1H, S2B). Collectively, these observations confirmed that organoid cell-fates can be fine-tuned via competing signalling pathways and organoid culture media should be carefully considered when modelling cell-types of interest (Figures 3A, S4C-F).

proCSC are enriched in CRC organoids and are transcriptionally similar to cells found in human and mouse CRC (Figure S2A). However, we demonstrated that proCSC are also present in WT epithelia and highly enriched in WT organoids cultured with WENR ligands. We therefore do not consider proCSC to be cancer stem cells. Rather than establishing an entirely new cancer-specific cell-fate, our study suggests that oncogenic mutations cell-intrinsically polarise cells to an extreme yet pre-existing proCSC state, while simultaneously disrupting cell-extrinsic regulation of plasticity – trapping cells as proCSC. These results describe cancer as a chronic, unidirectional shift in de-differentiation.

This study charts a continuous polarisation trajectory between revCSC and proCSC in colonic epithelia. In the healthy small intestine, revival stem cells have been demonstrated to act as multipotent stem cells that can be mobilised to replenish traditional $LGR5^+$ stem cells in response to tissue damage [7]. Small intestinal revival stem cells are found in the homeostatic small intestine *in vivo* [8, 33] and resemble an early 'foetal' stem cell-fate [24, 25]. Here we show that in colonic epithelia, revCSC are enriched by fibroblast-derived WNT3A and TGF-β via epithelial YAP, but only in the context of low PI3K and MAPK signalling. Our work and others now collectively suggest that fibroblasts are master regulators of revival stem cells in both the small intestine and colon.

Although revCSC are most easily accessible in WT epithelia, multiple studies have suggested revCSC also have an important role in CRC [9]. revCSC are candidates for early tumour initiating cells [8] and may confer WNT-inhibitor resistance in CRC [37]. A recent study in human CRC organoids also demonstrated that cancer cells can escape chemotherapy by adopting a slow-proliferating $Mex3a^+$ state driven by a low-EGF and high TGF-β culture environment [10]. Our results confirmed that TGF-β can induce revCSC-like cells in CRC organoids, but this process is rare (Figure S3C) and requires low PI3K signalling (Figure 5F). Moreover, we recently demonstrated that cancer associated fibroblasts (CAFs) can also induce a revCSC-like state in CRC patient-derived organoids (PDOs) that protects

CRC cells from chemotherapies including fluorouracil, oxaliplatin, and irinotecan [38]. In this model, CAF-chemoprotection can also be overcome by inhibiting YAP signalling – further demonstrating the central role of YAP in revCSC identity. However, CAF-chemoprotection is highly patient-specific, indicating only certain cell-states can be polarised to revCSC in CRC. Collectively, our results and others suggest fibroblast-induced revCSCs may represent an important 'drug-tolerant persister' (DTP) state in CRC. Given that targeting cell-plasticity is an emerging area of cancer therapies [39], future studies could target CRC DTP cells by combining YAP inhibitors (to block access to DTP revCSC) with standard chemotherapies (to kill proCSC).

In summary, through single-cell perturbation analysis of >1,000 organoid cultures, we charted a continuous landscape of cell-intrinsic and -extrinsic regulation of colonic stem cell polarisation. We found that colonic stem cell polarity is regulated by competing YAP and PI3K signalling flux, with stromal TGF-β pushing epithelia towards revCSC and CRC mutations trapping epithelia as proCSC. We conclude that cell-fate plasticity is a hallmark of colonic oncogenesis, and that cells can rapidly traverse the colonic differentiation landscape via combinations of oncogenic and stromal signalling.

Methods

Colonic Organoid Culture

Wild-type murine colonic organoids and CRC organoids carrying oncogenic mutations (*shApc* (A), $Kras^{G12D/+}$ (K), *shApc* and $Kras^{G12D/+}$ (AK), $Kras^{G12D/+}$ and $Trp53^{R172H/-}$ (KP), and *shApc*, $Kras^{G12D/+}$ and $Trp53^{R172H/-}$ (AKP)) were a kind gift from Lukas Dow (Cornell University) [40]. *shApc* was induced by Doxycycline treatment at 1 µg mL^{-1} and the efficiency of *Apc* knockdown was monitored with EGFP expression. Organoid base medium was made up of advanced DMEM/F-12 (Thermo 12634010) supplemented with 2 mM l-glutamine (Thermo 25030081), 1 mM N-acetyl-l-cysteine (Sigma A9165), 10 mM HEPES (Sigma H3375), 1× B-27 Supplement (Thermo 17504044), 1× N-2 Supplement (Thermo 17502048), and 1× HyClone Penicillin Streptomycin Solution (Fisher SV30010). Colonic organoids were cultured in organoid base medium further supplemented with 100 ng mL^{-1} murine WNT3A (mWNT3A, Peprotech 315-20), 50 ng mL^{-1} mEGF (Thermo PMG8041), 50 ng mL^{-1} mNoggin (Peprotech 250-38), 500 ng mL^{-1} mR-Spondin-1 (Peprotech 315-32), and 10 mM nicotinamide (Sigma N0636). WENR ligands were excluded from all experimental conditions throughout this study unless otherwise stated to ensure cell-cell signalling was not dominated by exogenous recombinant proteins.

For the WENR permutation experiment (Figures 2, 3, S3, and S4), colonic organoids were starved of mWNT3A, mEGF, mNoggin, and mR-Spondin-1 (WENR) for 6 h, split at a ratio of 1:3 (WT, A) or 1:6 (K, KP, AK, and AKP), and seeded as monocultures or fibroblast co-cultures at 5,000 fibroblasts per µL of Matrigel. The cultures were incubated with organoid base medium supplemented with 1× Insulin-Transferrin-Selenium-Sodium Pyruvate (ITS-A) (Thermo 51300044) and 10 mM nicotinamide (Sigma N0636) in addition to the combinations of mWNT3A (100 ng mL^{-1}), mEGF (50 ng mL^{-1}), mNoggin (50 ng mL^{-1}), and mR-Spondin-1 (500 ng mL^{-1}) as described in Figure 2. The cells were cultured for 48 h prior to TOB*is* MC analysis (see below).

For the WNT-EGF competition experiment (Figure S4C-F), WT colonic organoids were starved of mWNT3A, mEGF, mNoggin, and mR-Spondin-1 (WENR) for 6 h and split at a ratio of 1:3 and seeded as monocultures. WNT3A ranged from 0 to 100 ng mL^{-1} (0, 10, 20, 50, 100 ng mL^{-1}) and / or EGF ranged from 0 to 50 ng mL^{-1} (0, 10, 25, 40, 50 ng mL^{-1}) were added to the culture to capture their differential polarisation of revCSC and proCSC. The cells were cultured for 48 h prior to TOB*is* MC analysis (see below).

For the CellChat follow-up experiment (Figures 4D, S5D-F), colonic organoids were starved of mWNT3A, mEGF, mNoggin, and mR-Spondin-1 (WENR) for 6 h, split at a ratio of 1:3 (WT, A) or 1:6 (K, KP, AK, and AKP), and seeded as monocultures. The cells were incubated with organoid base medium supplemented with 1× ITS-A (Thermo 51300044), 10 mM nicotinamide (Sigma N0636), and the signalling ligands identified from the ligand-receptor analysis (Figure 4C): murine WNT5A (250 ng mL^{-1}, R&D Systems 645-WN-010/CF), murine SEMA3A (250 ng mL^{-1}, R&D Systems 5926-S3-025/CF), human TGF-β2 (1 ng mL^{-1}, BioLegend 583301), murine TGF-β1 (1 ng mL^{-1}, BioLegend 763102), murine IGF1 (100 ng mL^{-1}, Cell Guidance Systems GFM5-10), murine NRG1 (100 ng mL^{-1}, R&D Systems 9875-NR-050), murine EREG (500 ng mL^{-1}, R&D Systems 1068-EP-050/CF), and murine OPN (400 ng mL^{-1}, BioLegend 763604). Organoids treated with WNT3A (100 ng mL^{-1}) or EGF (50 ng mL^{-1}) were included as positive controls. The cells were cultured for 48 h prior to TOB*is* MC analysis (see below).

For the cue-signal-response MC array (Figures 5, S6), colonic organoids were starved of mWNT3A, mEGF, mNoggin, and mR-Spondin-1 (WENR) for 6 h, split at a ratio of 1:3 (WT, A) or 1:6 (K, AK), and seeded as monocultures. The cells were incubated with organoid base medium supplemented with 1× ITS-A, 10 mM nicotinamide, with or without signalling ligands: murine WNT3A (100 ng mL^{-1}), murine EREG (500 ng mL^{-1}), murine TGF-β1 (2 ng mL^{-1}). For each ligand condition, signalling inhibitors were added at the following concentrations: CA3 (YAP inhibitor, 2 µM, Sigma SML2647), CHIR99021 (GSK-3β inhibitor, 3 µM, Cell Guidance Systems SM13-1), ICG-001 (CBP/β-Catenin inhibitor, 2 µM, Cayman Chemical 16257), Dasatinib (SRC inhibitor, 50 nM, Cell Guidance Systems SM45-20), Trametinib (MEK inhibitor, 50 nM, Cayman Chemical 16292), GDC-

0941 (PI3K inhibitor, 1 µM, Selleck Chemical 50-851-6), PF-573228 (FAK inhibitor, 2.5 µM, Cayman Chemical CAY14924), and SIS3 (SMAD3 inhibitor, 3 µM, Cayman Chemical 15945). The cells were cultured for 48 h prior to TOB*is* MC analysis (see below).

Heterocellular Organoid Culture

The heterocellular organoid cultures were established as previously described [12]. Briefly, organoids were starved of mWNT3A, mEGF, mNoggin, and mR-Spondin-1 (WENR) for 6 h prior to the experiment and passaged at a ratio of 1:2.5; colonic fibroblasts (isolated, immortalised, and characterised in [12]) were seeded at 6,000 cells per µL for monoculture, 5,000 cells per µL for two-way co-cultures, and 4,000 cells per µL for three-way co-cultures; primary bone marrow-derived macrophages were seeded at 9,000 cells per µL for monoculture, 8,000 cells per µL for two-way co-cultures, and 7,000 cells per µL for three-way co-cultures. The cells were mixed in Matrigel and seeded at 7× 40 µL droplets per well in 6-well plates (for scRNA-seq) or 1× 50 µL droplet per well in 48-well plates (for TOB*is* MC). Unless otherwise specified, each microenvironment culture was maintained in WENR-free advanced DMEM/F-12 (Thermo 12634010) supplemented with 2 mM l-glutamine (Thermo 25030081), 1 mM N-acetyl-l-cysteine (Sigma A9165), 10 mM HEPES (Sigma H3375), 1× B-27 Supplement (Thermo 17504044), 1× N-2 Supplement (Thermo 17502048), 1× Insulin-Transferrin-Selenium-Sodium Pyruvate (ITS-A, Thermo 51300044) and 1× HyClone penicillin streptomycin solution (Fisher SV30010) for 48 h prior to TOB*is* MC analysis (see below).

scRNA-seq Data Acquisition

To prepare single-cell suspensions from the heterocellular organoid cultures, cells were removed from Matrigel using ice-cold PBS, collected with a benchtop centrifuge, and incubated with TrypLE™ Express Enzyme (Thermo 12604013) for 7 to 10 min at 37 ℃. The cells were then washed with ice-cold advanced DMEM/F-12 (Thermo 12634010) and filtered through a 35-µm cell strainer (Fisher 10585801). For FACS sorting, eBioscience™ Fixable Viability Dye eFluor™ 780 (FVD780, Thermo 65-0865-14) was used to label dead cells, while FITC anti-mouse CD66a (CEACAM1) antibody (Clone: MAb-CC1; BioLegend 134518) was used to stain epithelial cells, and APC anti-mouse CD45 antibody (Clone: BM8; BioLegend 123116) was used to stain macrophages. The gating of fibroblasts was based on their endogenous DsRed expression [12]. The collected cells were counted with a Countess II automated cell counter (Thermo Fisher) and examined for viability (samples with >90% viable cells were passed onto scRNA-seq library construction). To preserve RNA in the samples and to minimise technical variations, cells were fixed in ice-cold methanol immediately after counting as per the 10X Genomics instruction. For co-cultures, different cell-types were mixed at equal cell numbers prior to the fixation step. The methanol-fixed cells were stored at –20 ℃ for up to 2 weeks before they were rehydrated and processed using the 10X Genomics Chromium Controller. scRNA-seq libraries were generated with the 10X Genomics Chromium Next GEM Single Cell 3' Reagent Kits v3.1 (Dual Index) and sequenced with the Illumina NovaSeq 6000 System (2× 150 bp paired-end reads), aiming at 60,000 read pairs per cell and 2,000 cells per cell-type per sample.

scRNA-seq Data Processing

Raw binary base call (BCL) sequence files were converted to FASTQ files and processed with the 10X Genomics Cell Ranger pipeline version 5.0.1. The FASTQ files were then aligned to a custom GRCm38 reference genome containing the sequences of *DsRed* and *eGFP* transgenes present in fibroblasts and organoids respectively, generating pre-filtered feature-barcode matrices.

The gene count matrices were analysed with the R package *Seurat* version 4.0.4 [16]. The analysis pipeline encompasses quality control, data normalisation, data integration, dimensionality reduction, cell clustering, and analysis of differential gene expression. Genes found in less than 4 cells were removed during QC and only cells with > 600 unique genes identified were kept for downstream analysis. The total number of detected sequences typically ranged from 1,200 to 80,000 per cell, and the actual values were manually determined based on cell-type composition and sequencing depth. For the integrated epithelial object in Figure 1D, an additional filtering step was performed to remove cells with undetectable expression for any one of the bona fide pan-epithelial genes *Epcam*, *Krt8*, *Krt18*, *Krt19*, *Cldn7*. Cell-cycle regression was performed using the *sctransform* function. Log-normalised gene expression values (RNA assay) were used for downstream analysis if not otherwise stated.

Dataset integration was performed using Seurat's reciprocal PCA (RPCA) implementation [16] (k.anchor=12) as it has been optimised to handle large datasets. The integrated object in Figure 1B was computed using all cells from the 20 conditions shown in Figure 1A (integrated object limited to 2,000 genes across 58,726 cells). The integrated object in Figure 1D was computed using just the epithelial cells from all conditions (4,000 genes, 29,452 cells).

For dimensionality reduction (DR), the first 50 principal components (PC) was computed from the integrated assays to generate 2-dimensional PHATE embeddings with default parameters (Table S2). PHATE was chosen as the standard DR method for the study due to its capacity to capture the global structure in biological systems with important developmental trajectories [17].

Cell clustering was computed using the Leiden algorithm on the *k*NN graph generated from the integrated epithe-

lial dataset (first 48 PCs), at a series of resolutions ranging from 0.2 to 0.8. The final cluster annotations were retrospectively defined by common cell-type marker expression, inter-cluster relationships on a multi-resolution clustering tree [41], and cross-condition differential abundance behaviours (see below). Cells from outlier clusters (totalling less than 1% of all epithelial cells) were excluded from the downstream analysis.

Differentially expressed (DE) genes between clusters, conditions, and cell neighbourhoods were identified using Wilcoxon rank-sum test implemented by Seurat's *FindAllMarkers* and *FindMarkers* functions.

scRNA-seq Data Analysis

To generate the EMD-PCAs in Figure 1C, log-normalised gene expression data of all cells of a particular cell-type (epithelial cells, fibroblasts, or macrophages) were exported from the integrated object. EMD scores for the top 6,000 variable genes of each condition were calculated with CyGNAL [42] using the WT monoculture control as the reference.

Differentially abundant (DA) cell neighbourhoods were identified using the R package *MiloR* [23], which enabled the detection of enrichment and depletion of cell clusters caused by microenvironmental and/or genotypical perturbations. Given that $CD34^{hi}$ and $CD34^{lo}$ fibroblasts do not differentially regulate epithelial cells (Figure S1B), all samples of WT organoid+fibroblast co-cultures were grouped and considered replicates of the query condition regardless of the CD34 status of the fibroblasts, with the DA test threshold set at 5% SpatialFDR (Figure 1F). Similarly, AK and AKP organoid monocultures were grouped due to their similar DE and DA behaviour (Figure 1G). The DA overview dot plot in Figure 1H was generated by comparing the 17 conditions against the WT monoculture control (2× replicates).

Heatmaps of selected marker genes were generated with the R package *ComplexHeatmap* [43] across the manuscript. Gene lists in Figures 1I and Figure S1B were curated from previously reported markers for colonic epithelial subpopulations and DE genes detected between epithelial clusters, conditions, and DA neighbourhoods within this study. Gene lists in For S1B-D represent DE genes between conditions.

The *UCell* [44] method was used to generate the correlation matrix between gene signatures in existing literature and cell clusters identified within this study (Figure S2A). Gene lists for different intestinal stem cell-states were compiled from public datasets, together with transcriptional targets of key signalling pathways encoding the different stem cell-states (Table S3). These gene lists were compared with the curated gene signatures for proliferation, CSC, revCSC, and proCSC cell-states in Figure 1I, as well as the top DE genes for each stem cluster (adjusted p-value < 0.01, log2FC > 0.25, top 24 genes with the greatest positive log2FC values) (Table S3). *UCell* scores for each gene set were calculated using Log-normalised gene expression values and z-scored to allow cross-signature comparison. Pearson correlations were computed between the scores on all cells of stem and TA clusters and then visualised as a correlation heatmap, grouped via complete linkage hierarchical clustering, only showing significant correlations (conf.level = 0.95).

Leveraging the concept that cells with a higher potency should have a higher signalling entropy [45], the pluripotency values for epithelial cells across the different clusters were estimated using the R package *SCENT* [21]. Signalling entropy scores for all epithelial cells were computed with the CCAT (correlation of connectome and transcriptome) approximation method using a murinised version of the built-in *net17Jan16* Protein-Protein interaction network.

For RNA velocity analysis, loom files were generated from Cell Ranger's output using the Python package *velocyto* [46] (reference genome: GRCm38, repeat mask assembly: GRCm38/mm10, track: RepeatMasker). RNA velocity was analysed with the Python package *scVelo* [22] using default parameters unless otherwise specified (Table S2). Metadata and PHATE embedding coordinates were exported from the relevant Seurat objects to filter and annotate anndata objects generated from the loom files made by velocyto. Moments for the velocity estimation were calculated using the first 50 PCs and 30 neighbours from the anndata objects. RNA velocities were computed with the *recover_dynamics* function using the dynamical model of transcriptional dynamics with default parameters. The velocity stream embedding (Figure 1E) was computed using the integrated object containing epithelial cells from all conditions. The RNA velocity vector lengths, an estimate of a cell's differentiation rate, were computed using cells solely from the 4 conditions shown in Figure S2B-D. The quantitative comparison in Figure S2D was performed using the Games-Howell pairwise test wrapper from the R package *statsExpressions* [47]. All conditions were compared against the WT monoculture control and all *p*-values have been corrected for multiplicity with the Holm method.

Ligand-receptor expression analysis was performed using the R package CellChat [28], where stromal-epithelial signalling was analysed across 4 different organoid genotypes (WT, A, AK, and AKP). Epithelial cells were annotated with the clusters previously identified (Figure 1D), while the fibroblasts were grouped as a single cluster. A merged CellChat object was generated to compare relative communication probability of fibroblast-to-epithelia signalling across the genotypes. Significant ligand-receptor pairs were identified based on CellChat's murine cell communication database. Plots displaying aggregate outgoing and incoming communication probability (Figure 4A) were generated with the *netAnalysis_signalingRole_scatter* function. Detected communication at the pathway and interaction level was accessed

with the *subsetCommunication* function and probabilities were z-score normalised to allow for cross-pathway or cross-interaction comparison. The results were visualised with ComplexHeatmap in Figure 4B-C, the rows of which were manually ordered based on hierarchical clustering and grouped based on the nature of the interaction. Gene expression of the ligand-receptor pairs identified above was visualised using Seurat's *Dotplot* function in Figure S5A. *UCell* scores for ligand and receptor genes were calculated for fibroblasts and epithelial cells respectively. Games-Howell pairwise test was performed using the R package *statsExpressions* and all *p*-values have been corrected for multiplicity with the Holm method.

TOB*is* MC

TOB*is* MC of organoid cultures was performed as previously described [13]. Briefly, the cultures were incubated with 25 µM ^{127}5-iodo-2'-deoxyuridine (^{127}IdU) for 30 min to label S-phase cells, treated with a cocktail of protease (Sigma P8340) and phosphatase inhibitors (Sigma 4906845001) to protect protein and phosphorylation epitopes, and fixed with 4% (w/v) PFA for 1 h at 37 ℃. The cells were washed twice with PBS, incubated in 250 nM $^{194/8}$cisplatin (Fluidigm 201194/8) for 10 min to stain dead cells, and washed twice with PBS to remove residual cisplatin. TOB*is* barcodes were added to the cells and incubated overnight at 4 ℃. The following day, unbound barcodes were quenched with reduced glutathione (Sigma G6529) and washed from the cultures. TOB*is*-barcoded organoids from each condition were removed from Matrigel in a freshly prepared dissociation buffer containing 0.5 mg mL^{-1} Dispase II (Thermo 17105041), 0.2 mg mL^{-1} Collagenase IV (Thermo 17104019) and 0.2 mg mL^{-1} DNase I (Sigma DN25), pooled into a single master tube and dissociated into single cells with a gentleMACS Octo Dissociator (Miltenyi 130-096-427). Following dissociation, the cells were washed, filtered, and stained for extracellular epitopes with rare earth metal-labelled antibodies (Table S1). The cells were then permeabilised with 0.1% (v/v) Triton X-100 followed by 50% (v/v) methanol. Once permeabilised, the cells were stained with a panel of metal antibodies against intracellular proteins and PTMs (Table S1). For each cell, we measured cell-type markers (epithelia: CEACAM-1, Pan-cytokeratin (Pan-CK), GFP; fibroblasts: PDPN, RFP, mCherry), epithelial differentiation markers identified by scRNA-seq (CLU, CD44, SOX9, SURVIVIN, LRIG1, EPHB2, C-MYC, and FABP2), cell-state markers (pRB [S807/S811], IdU, pHH3 [S28], Cyclin B1, and cCaspase3 [D175] [13]), and >20 PTMs spanning multiple cell-signalling pathways. The cells were washed and incubated in DNA intercalator $^{191/193}$Ir (Fluidigm 201192A) overnight before MC single-cell data acquisition and analysis.

MC Data Acquisition and Analysis

TOB*is* MC data were acquired and analysed as previously described [13]. For Fluidigm Helios acquisitions, stained cells were washed into Maxpar Water (Fluidigm 201069) containing 2 mM EDTA, diluted to 0.8–1.2$\times10^6$ cells mL^{-1} and spiked with EQ Four Element Calibration Beads (Fluidigm 201078). The cells were then loaded into a Super Sampler (Victorian Airships). For CyTOF XT acquisitions, stained cells were wash into Maxpar Cell Acquisition Solution Plus (Fluidigm 201244) containing 2 mM EDTA, diluted to 0.8–1.2$\times10^6$ cells mL^{-1} and spiked with EQ™ Six Element Calibration Beads (Fluidigm 201245).

After data acquisition, raw MC data were normalised and exported as standard FCS file(s). Multiplexed TOB*is* experiments were debarcoded into individual conditions (https://github.com/zunderlab/single-cell-debarcoder), imported into Cytobank (http://www.cytobank.org/), and gated with Gaussian parameters, DNA/cisplatin, and cell-type markers to remove debris, identify live cells, and remove doublets respectively. The fully gated datasets were further processed with our MC data analysis pipeline, CyGNAL (https://github.com/TAPE-Lab/CyGNAL) [42]. Earth mover's distance (EMD) [48] was used to quantify node intensity of each marker. Unless otherwise specified, EMD scores were calculated with WT untreated controls (concatenated replicates) as the reference.

PHATE [17] embeddings were calculated with raw/z-scored EMD scores or arcsinh-transformed single-cell MC data using the python package *phate* (https://github.com/KrishnaswamyLab/PHATE) with parameters specified in Table S2. EMD heatmaps were generated with the R package ComplexHeatmap [43] and further annotated in OmniGraffle Professional across the manuscript. For the WENR permutation experiment (Figures 2, 3, S3, and S4), EMD scores for revCSC and proCSC markers (Figure S3A), percentages of S-phase cells (Figure S3B) and CLU$^+$ cells (Figure S3C) were plotted and analysed with GraphPad Prism 7 (ordinary one-way ANOVA with Holm-Šídák's multiple comparisons test for Figure S3A, unpaired two-tailed *t*-tests for Figure S3B-C). For the cue-signal-response perturbation array (Figures 5, S6), EMD scores for CLU and LRIG1 were calculated for selected conditions and analysed with GraphPad Prism 7 (ordinary one-way ANOVA with Holm-Šídák's multiple comparisons or unpaired two-tailed *t*-tests for Figure 5G, H).

Force-directed Scaffold Maps [27] (Figure 3C) were constructed using the R package Scaffold (https://github.com/nolanlab/scaffold). Landmark populations (WT, A, K, KP, AK, AKP organoid monocultures) were manually gated and exported from Cytobank with all data arcsinh transformed (cofactor = 5). The parameters used in the Scaffold analysis were specified in Table S2.

The Boolean logic models of CSC regulation by WENR ligands (Figure 3D) were compiled in OmniGraffle Professional.

The relative stemness (RS) score (Figure 5F) was generated by calculating ratios between arcsinh-transformed LRIG1 and CLU MC measurements for single cells, followed by $\log_2$ normalisation, and then summarised at the replicate and condition level. RS score heatmap (Figure S6B) was generated with the R package ComplexHeatmap [43] and further annotated in OmniGraffle Professional.

Valley-Ridge (VR) Score

The VR score was defined as the weighted sum of the Valley (weight = 0.9) and the Ridge (weight = 0.1) components, and was computed on a per sample and per cluster basis (Figure S7A). The Valley component equals the median CCAT value of each sample-cluster combination. To calculate the Ridge component, the inverse of the velocities was first computed and scaled to a range between 0 and 1. A cell centrality distance was then calculated for cells in each cluster by first building a *k*NN graph of a cluster's cells from the PHATE embeddings (Table S2), followed by the calculation of a distance matrix using graphtool's *shortest_path* function [49]. The median distance for each cell to all other cells was then calculated, whereby cells with the lowest distance would be at a cluster's centre whilst those with the highest distance would be at the cluster periphery. To allow inter-cluster comparisons, outliers with a distance over Q_{99} were set to the median distance value before scaling to (0,1). Finally, the Ridge component was computed per sample-cluster as the product of the median scaled inverse velocities and the cell's scaled centrality distance (Figure S7A).

This definition of the VR score allows the CCAT-driven Valley component to be the driving force for sculpting the landscape and the velocity-driven Ridge component to predominately define the barriers around clusters – producing a tarn-like effect symbolising a state of trapped cells. In principle, any other dimensionality reduction technique can be used in place of PHATE [50], and the Valley/Ridge component can be computed using other metrics underpinning pluripotency and cell-fate transition. The Ridge component can also be calculated with a distance-free approach such as α-shapes [51]. Finally, the VR scores can be computed on a per cell or neighbourhood basis, which will increase landscape resolution and liberate the method from constraints of cluster definitions (at the expense of increased noise).

Waddington-like Landscape

To generate the Waddington-like landscapes in Figure 6A, we combine the ability of PHATE to capture the global structure of single-cell data with the VR score (described above) (Figure S7A-B).

Waddington-like landscapes can be visualised directly in Python (Figure S7B, C). Briefly, a low dimensional 34x30 mesh grid was generated from the PHATE embeddings, and a 3D surface was rendered by projecting VR scores onto the grid using the radial basis function interpolation from scipy [52] (Table S2). The surface of the landscape was coloured by VR scores and a scatter plot was overlaid where the elevation of each cell was defined as the weighted sum of its VR score (weight = 0.9), CCAT value (weight = 0.1), and a constant factor of 0.012 (weight = 1). This added a level of controlled noise to the scatter plot while ensuring most cells remain above the interpolated surface (Figure S7C).

These landscapes can also be visualised in SideFX Houdini 19.5 (http://www.sidefx.com) and rendered using Maxon Redshift 3.5 (http://www.redshift3d.com) (Figures 6A, S7B). VR scores and scRNA-seq metadata were imported and points were positioned in z- and x-axes according to their PHATE scores. This PHATE distribution was then transformed in the y-axis according to each cell's VR score. The PHATE-transformed 2D distribution was used as a deformation lattice to influence nearby points on a polygonal grid, and its difference from the VR-transformed 3D distribution was used to drive deformation of this polygonal grid into a Waddington-like landscape. The VR-transformed data was then projected back onto the Waddington-like landscape to avoid intersections between positions of data points and landscape topology. A video tutorial to visualise Waddington-like embeddings using Houdini is available at: https://entagma.com/houdini-tutorial-waddington-landscape/.

Data Availability

Raw scRNA-seq data and BioSample metadata have been deposited at Sequence Read Archive (SRA) (https://www.ncbi.nlm.nih.gov/bioproject/PRJNA883610). Raw and processed MC data are available as a Community Cytobank project (https://community.cytobank.org/cytobank/experiments#project-id=1460). Aligned scRNA-seq count matrices, spliced/unspliced RNA count matrices, integrated Seurat objects, integrated MC dataframes, and Houdini project files can be accessed at Zenodo (https://doi.org/10.5281/zenodo.7586958). All analysis scripts to reproduce figure plots together with a notebook explaining pre-processing and QC steps for scRNA-seq analysis are available at GitHub (https://github.com/TAPE-Lab/Qin-CardosoRodriguez-et-al).

Acknowledgements

We are extremely grateful to L. Dow for sharing murine colonic organoids and S. Acton for providing murine tissue for fibroblast and macrophage isolation. We thank Y. Guo and the UCL CI Flow-Core for CyTOF support. We thank E. Sahai, V. Li, S. Acton, and members of the Tape Lab for their constructive critique of the manuscript. This work was supported by Cancer Research UK (C60693/A23783), the Cancer Research UK City of London Centre (C7893/A26233), the UCLH

Biomedical Research Centre (BRC422), and the UKRI Medical Research Council (MR/T028270/1).

Author Contributions

X.Q. designed the study, performed organoid experiments, generated scRNA-seq and TOB*is* mass cytometry data, analysed mass cytometry data, and wrote the paper. F.C.R. analysed scRNA-seq data, developed the VR score, and wrote the paper. J.S. developed TOB*is* barcodes and conjugated rare-earth metal antibodies. P.V. provided organoid culture support. J.C rendered Waddington-like landscapes. C.J.T. designed the study, analysed the data, and wrote the paper.

Co-authors reserve the right to rearrange authorship positions on their CVs.

References

[1] N. Sphyris, M.C. Hodder, and O.J. Sansom. Subversion of Niche-Signalling Pathways in Colorectal Cancer: What Makes and Breaks the Intestinal Stem Cell. Cancers (Basel) *13*.5 (2021), 1000.

[2] H. Gehart and H. Clevers. Tales from the Crypt: New Insights into Intestinal Stem Cells. Nat Rev Gastroenterol Hepatol *16*.1 (2019), 19–34.

[3] H. Clevers. The Intestinal Crypt, A Prototype Stem Cell Compartment. Cell *154*.2 (2013), 274–284.

[4] M. van de Wetering et al. The Beta-Catenin/TCF-4 Complex Imposes a Crypt Progenitor Phenotype on Colorectal Cancer Cells. Cell *111*.2 (2002), 241–250.

[5] J. Beumer and H. Clevers. Cell Fate Specification and Differentiation in the Adult Mammalian Intestine. Nat Rev Mol Cell Biol *22*.1 (2021), 39–53.

[6] Y.M. Nusse et al. Parasitic Helminths Induce Fetal-like Reversion in the Intestinal Stem Cell Niche. Nature *559*.7712 (2018), 109–113.

[7] A. Ayyaz et al. Single-Cell Transcriptomes of the Regenerating Intestine Reveal a Revival Stem Cell. Nature *569*.7754 (2019), 121–125.

[8] M. Roulis et al. Paracrine Orchestration of Intestinal Tumorigenesis by a Mesenchymal Niche. Nature *580*.7804 (2020), 524–529.

[9] E.G. Vasquez et al. Dynamic and Adaptive Cancer Stem Cell Population Admixture in Colorectal Neoplasia. Cell Stem Cell *29*.8 (2022), 1213–1228.e8.

[10] A. Álvarez-Varela et al. Mex3a Marks Drug-Tolerant Persister Colorectal Cancer Cells That Mediate Relapse after Chemotherapy. Nat Cancer *3*.9 (2022), 1052–1070.

[11] C.J. Tape. The Heterocellular Emergence of Colorectal Cancer. Trends Cancer *3*.2 (2017), 79–88.

[12] X. Qin et al. Cell-Type-Specific Signaling Networks in Heterocellular Organoids. Nat Methods *17*.3 (2020), 335–342.

[13] J. Sufi et al. Multiplexed Single-Cell Analysis of Organoid Signaling Networks. Nat Protoc *16*.10 (2021), 4897–4918.

[14] M. Roulis and R.A. Flavell. Fibroblasts and Myofibroblasts of the Intestinal Lamina Propria in Physiology and Disease. Differentiation *92*.3 (2016), 116–131.

[15] R.A. Isidro and C.B. Appleyard. Colonic Macrophage Polarization in Homeostasis, Inflammation, and Cancer. Am J Physiol Gastrointest Liver Physiol *311*.1 (2016), G59–G73.

[16] Y. Hao et al. Integrated Analysis of Multimodal Single-Cell Data. Cell *184*.13 (2021), 3573–3587.e29.

[17] K.R. Moon et al. Visualizing Structure and Transitions in High-Dimensional Biological Data. Nat Biotechnol *37*.12 (2019), 1482–1492.

[18] M.M. Mahe et al. Establishment of Gastrointestinal Epithelial Organoids. Curr Protoc Mouse Biol *3*.4 (2013), 217–240.

[19] I. Stzepourginski et al. CD34+ Mesenchymal Cells Are a Major Component of the Intestinal Stem Cells Niche at Homeostasis and after Injury. PNAS *114*.4 (2017), E506–E513.

[20] O.N. Karpus et al. Colonic CD90+ Crypt Fibroblasts Secrete Semaphorins to Support Epithelial Growth. Cell Rep *26*.13 (2019), 3698–3708.e5.

[21] A.E. Teschendorff and T. Enver. Single-Cell Entropy for Accurate Estimation of Differentiation Potency from a Cell's Transcriptome. Nat Commun *8* (2017), 15599.

[22] V. Bergen et al. Generalizing RNA Velocity to Transient Cell States through Dynamical Modeling. Nat Biotechnol *38*.12 (2020), 1408–1414.

[23] E. Dann et al. Differential Abundance Testing on Single-Cell Data Using k-Nearest Neighbor Graphs. Nat Biotechnol *40*.2 (2022), 245–253.

[24] R.C. Mustata et al. Identification of Lgr5-independent Spheroid-Generating Progenitors of the Mouse Fetal Intestinal Epithelium. Cell Rep *5*.2 (2013), 421–432.

[25] S. Yui et al. YAP/TAZ-Dependent Reprogramming of Colonic Epithelium Links ECM Remodeling to Tissue Regeneration. Cell Stem Cell *22*.1 (2018), 35–49.e7.

[26] E.R. Zunder et al. Palladium-Based Mass Tag Cell Barcoding with a Doublet-Filtering Scheme and Single-Cell Deconvolution Algorithm. Nat Protoc *10*.2 (2015), 316–333.

[27] M.H. Spitzer et al. An Interactive Reference Framework for Modeling a Dynamic Immune System. Science *349*.6244 (2015), 1259425.

[28] S. Jin et al. Inference and Analysis of Cell-Cell Communication Using CellChat. Nat Commun *12*.1 (2021), 1088.

[29] D. Dimitrov et al. Comparison of Methods and Resources for Cell-Cell Communication Inference

from Single-Cell RNA-Seq Data. Nat Commun *13*.1 (2022), 3224.

[30] C. Waddington. The Strategy of the Genes (London, UK: George Allen & Unwin, 1957).

[31] X. Qin and C.J. Tape. Deciphering Organoids: High-Dimensional Analysis of Biomimetic Cultures. Trends Biotechnol *39*.8 (2021), 774–787.

[32] J.S. Fleck et al. Inferring and Perturbing Cell Fate Regulomes in Human Brain Organoids. Nature (2022), 1–8.

[33] J. Bues et al. Deterministic scRNA-seq Captures Variation in Intestinal Crypt and Organoid Composition. Nat Methods *19*.3 (2022), 323–330.

[34] E. Sahai et al. A Framework for Advancing Our Understanding of Cancer-Associated Fibroblasts. Nat Rev Cancer *20*.3 (2020), 174–186.

[35] A. Merenda, N. Fenderico, and M.M. Maurice. Wnt Signaling in 3D: Recent Advances in the Applications of Intestinal Organoids. Trends Cell Biol *30*.1 (2020), 60–73.

[36] T. Sato et al. Long-Term Expansion of Epithelial Organoids from Human Colon, Adenoma, Adenocarcinoma, and Barrett's Epithelium. Gastroenterology *141*.5 (2011), 1762–1772.

[37] T. Han et al. Lineage Reversion Drives WNT Independence in Intestinal Cancer. Cancer Discov *10*.10 (2020), 1590–1609.

[38] M.R. Zapatero et al. *Trellis Single-Cell Screening Reveals Stromal Regulation of Patient-Derived Organoid Drug Responses*. Preprint. 2022.

[39] D.B. Burkhardt et al. Mapping Phenotypic Plasticity upon the Cancer Cell State Landscape Using Manifold Learning. Cancer Discov *12*.8 (2022), 1847–1859.

[40] L.E. Dow et al. Apc Restoration Promotes Cellular Differentiation and Reestablishes Crypt Homeostasis in Colorectal Cancer. Cell *161*.7 (2015), 1539–1552.

[41] L. Zappia and A. Oshlack. Clustering Trees: A Visualization for Evaluating Clusterings at Multiple Resolutions. GigaScience *7*.7 (2018), giy083.

[42] F. Cardoso. *TAPE-Lab/CyGNAL: V0.2.1*. Zenodo. 2021.

[43] Z. Gu, R. Eils, and M. Schlesner. Complex Heatmaps Reveal Patterns and Correlations in Multidimensional Genomic Data. Bioinformatics *32*.18 (2016), 2847–2849.

[44] M. Andreatta and S.J. Carmona. UCell: Robust and Scalable Single-Cell Gene Signature Scoring. Comput Struct Biotechnol J *19* (2021), 3796–3798.

[45] A.E. Teschendorff, P. Sollich, and R. Kuehn. Signalling Entropy: A Novel Network-Theoretical Framework for Systems Analysis and Interpretation of Functional Omic Data. Methods *67*.3 (2014), 282–293.

[46] G. La Manno et al. RNA Velocity of Single Cells. Nature *560*.7719 (2018), 494–498.

[47] I. Patil. statsExpressions: R Package for Tidy Dataframes and Expressions with Statistical Details. JOSS *6*.61 (2021), 3236.

[48] D.Y. Orlova et al. Earth Mover's Distance (EMD): A True Metric for Comparing Biomarker Expression Levels in Cell Populations. PLoS ONE *11*.3 (2016), e0151859.

[49] T.P. Peixoto. *The Graph-Tool Python Library*. 2017.

[50] W. Chen et al. Single-Cell Landscape in Mammary Epithelium Reveals Bipotent-like Cells Associated with Breast Cancer Risk and Outcome. Commun Biol *2*.1 (2019), 306.

[51] K. Bellock, N. Godber, and P. Kahn. *Bellockk/Alphashape: V1.3.1 Release*. Zenodo. 2021.

[52] P. Virtanen et al. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. Nat Methods *17*.3 (2020), 261–272.

Supplementary Information

Supplementary Figures

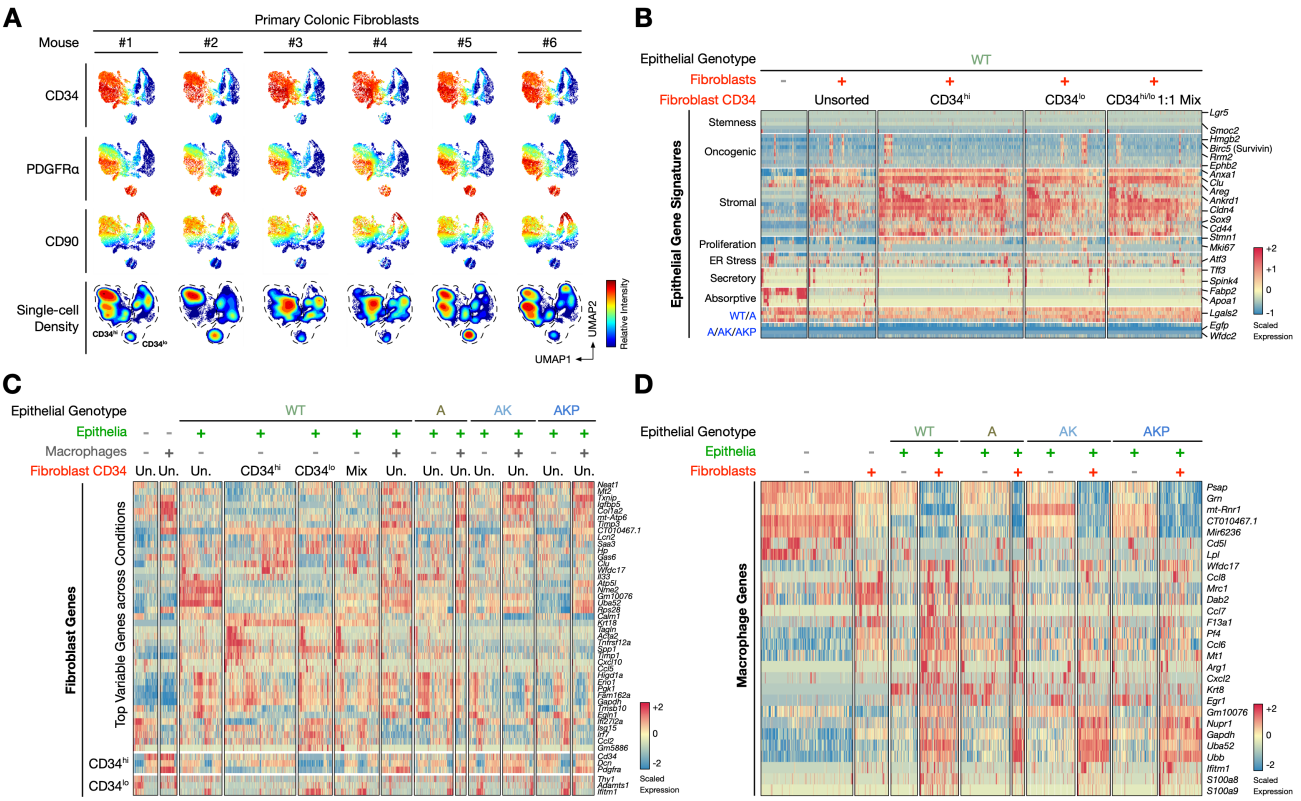


Figure S1. Single-cell Characterisation of the Heterocellular Organoid Model. Related to Figure 1. A) Mass cytometry (MC) analysis of primary murine colon fibroblasts showing stromal markers CD34, PDGFRα, and CD90. **B)** scRNA-seq analysis of WT colonic organoids co-cultured with unsorted, $CD34^{hi}$, $CD34^{lo}$, and a 1:1 mix of $CD34^{hi}$:$CD34^{hi}$ colonic fibroblasts. **C)** Differential gene expression analysis of fibroblasts regulated by epithelial organoids and macrophages. **D)** Differential gene expression analysis of macrophages regulated by epithelial organoids and fibroblasts.

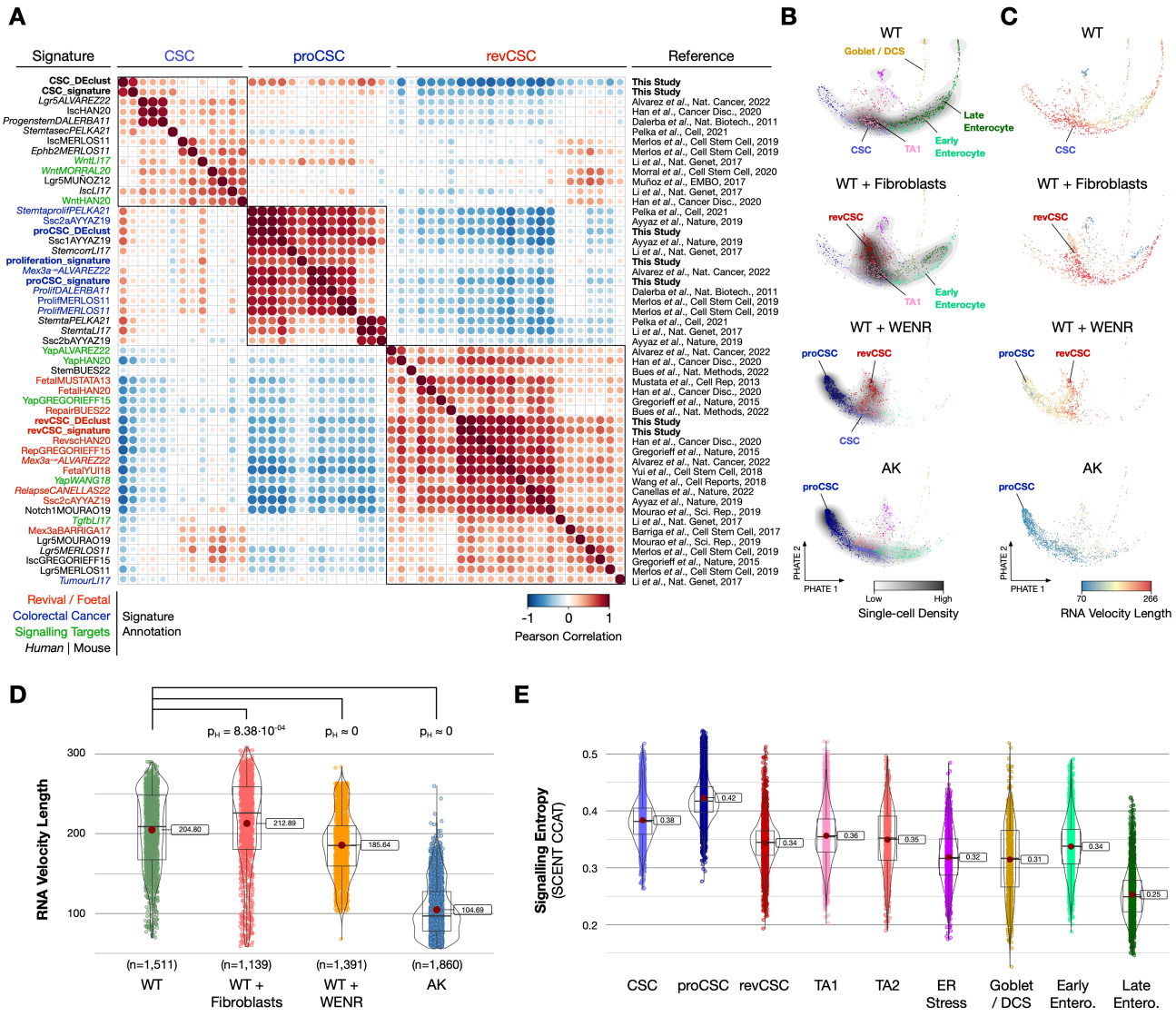


Figure S2. Epithelial Stem Cell Signatures. Related to Figure 1. A) Comparison of gene signatures of CSC, proCSC, and revCSC identified in this study with published stem cell signatures. **B)** Single-cell PHATE embeddings of epithelial cells from WT, WT+Fibroblasts, WT+WENR, and AK organoids coloured by cluster and overlaid with single-cell density. **C)** Single-cell PHATE embeddings coloured by RNA velocity vector lengths. **D)** RNA velocity vector lengths of organoid conditions (Games-Howell pairwise test with Holm-adjusted *p*-values). **E)** CCAT scores of epithelial clusters. CSC, colonic stem cell. proCSC, hyper-proliferative CSC. revCSC, revival CSC. TA, transit amplifying cell. DCS, deep crypt secretory cell. Entero., enterocyte. Boxplots show min/max and quartiles. Red dot marks the mean value.

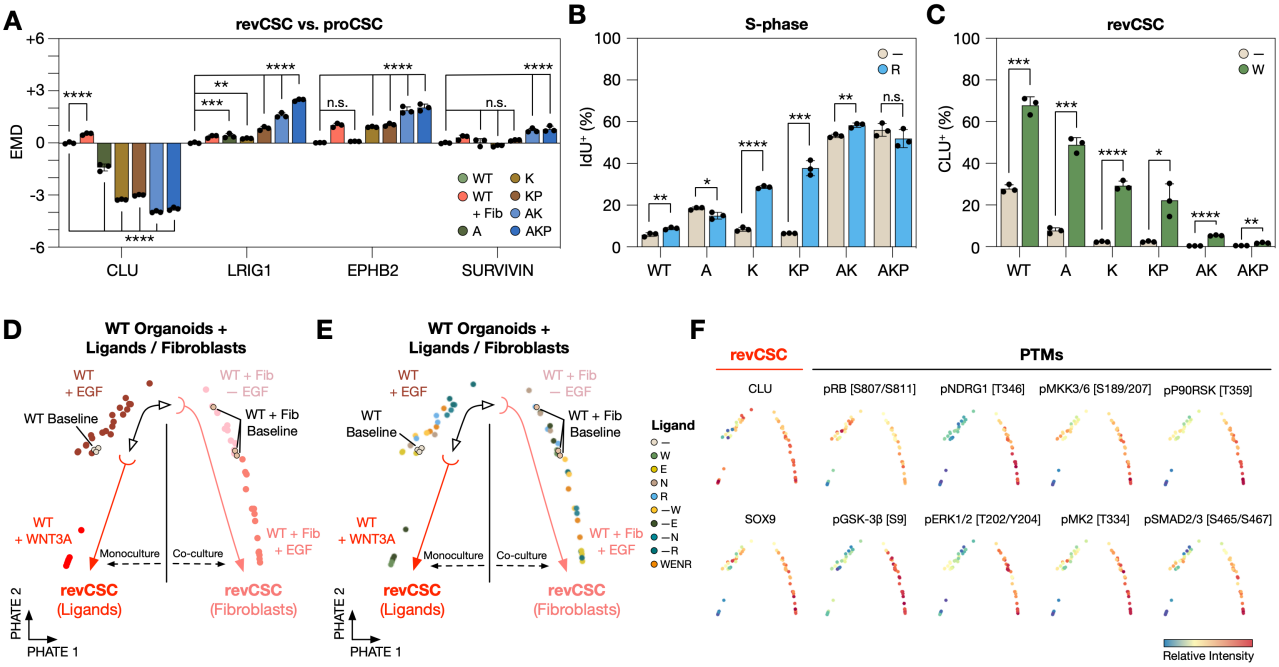


Figure S3. Epithelial Regulation by Organoid Genotypes, WENR Ligands, and Fibroblasts. Related to Figure 2. A) EMD scores for CLU, LRIG1, EPHB2, and SURVIVIN across organoid culture conditions (**, $p < 0.01$; ***, $p < 0.001$; ****, $p < 0.0001$. Ordinary one-way ANOVA with Holm-Šídák's multiple comparisons test). Error bars represent standard deviation (SD). **B)** Percentage of S-phase cells in organoids cultured with or without R-Spondin-1 across genotypes (***, $p < 0.001$; ****, $p < 0.0001$. Unpaired two-tailed *t*-test). Error bars represent SD. **C)** Percentage of revCSC in organoids cultured with or without WNT3A across genotypes (***, $p < 0.001$. Unpaired two-tailed *t*-test). Error bars represent SD. **D-F)** EMD-PHATE of WT organoids cultured with or without fibroblasts and WENR ligands coloured by microenvironment, ligands, and EMD scores for selected markers. One dot = one condition. revCSC, revival colonic stem cell. proCSC, hyper-proliferative colonic stem cell.

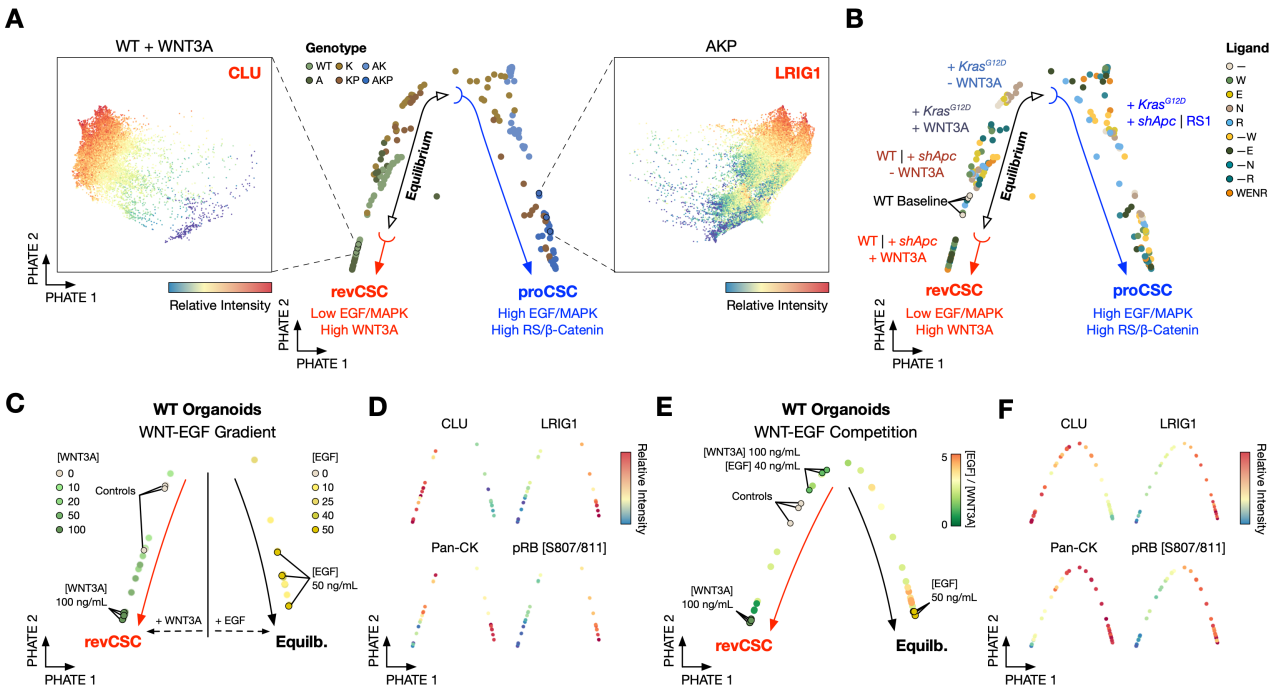


Figure S4. Stepwise Regulation of Epithelial Differentiation. Related to Figure 3. A) EMD-PHATE of organoid monocultures +/-WENR ligands coloured by organoid genotypes. Inserts: single-cell PHATE overlaid with arcsinh-transformed measurements of CLU or LRIG1 for WT+WNT3A versus AKP monoculture respectively. **B)** EMD-PHATE of organoid monocultures +/-WENR ligands annotated by organoid genotype and culture conditions. **C)** EMD-PHATE of WT organoids cultured with a gradient of either WNT3A or EGF coloured by WNT3A or EGF concentrations (ng mL^{-1}). **D)** The PHATE embedding in **C)** coloured by EMD scores for selected markers. **E)** EMD-PHATE of WT organoids cultured with varying combinations of WNT3A and EGF coloured by the ratio between EGF and WNT3A concentrations. **F)** The PHATE embedding in **E)** coloured by EMD scores for selected markers. Equilb., Equilibrium.

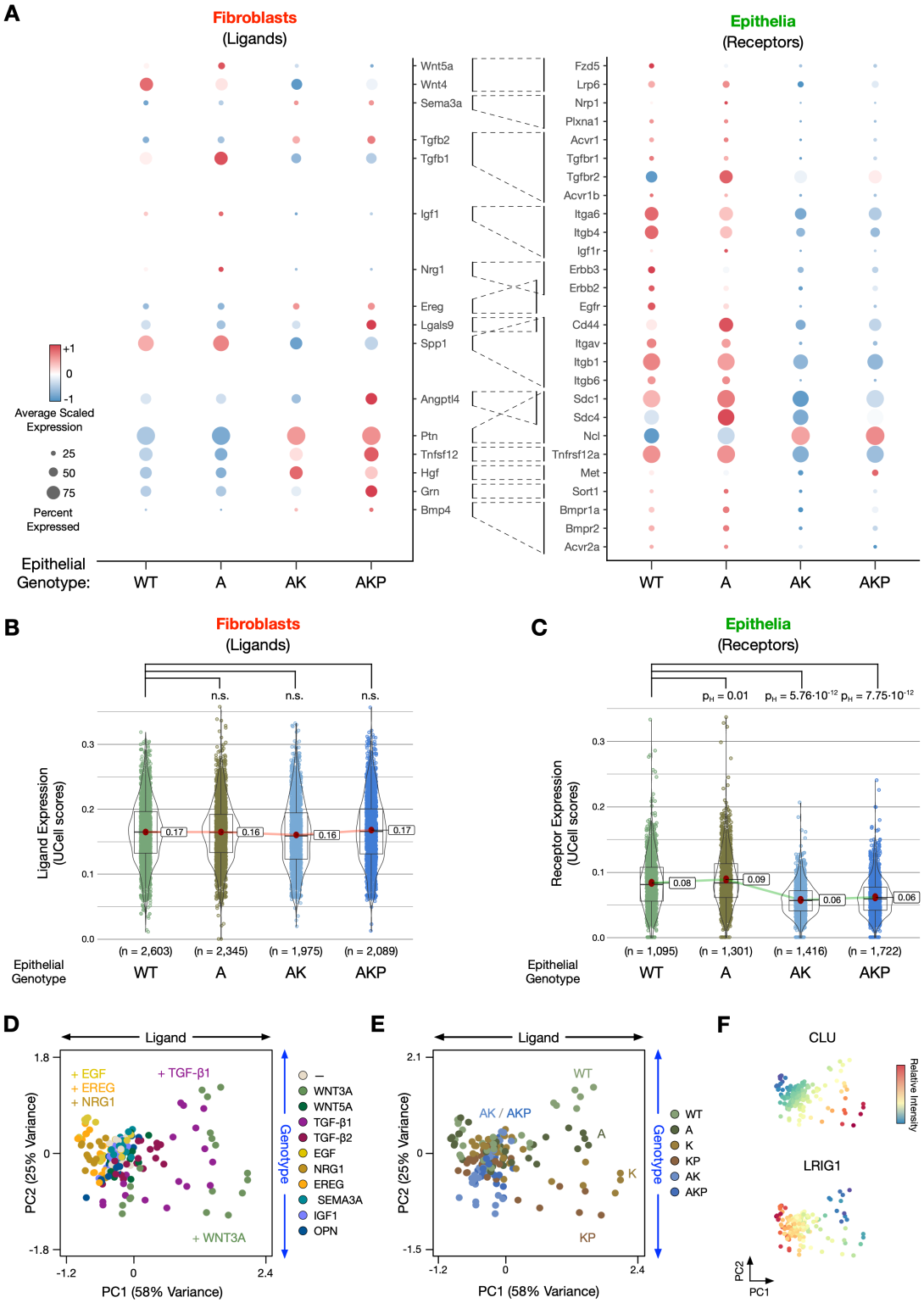


Figure S5. Ligand-receptor Expression Analysis. Related to Figure 4. A) Average scaled expression of ligands (expressed by fibroblasts) and receptors (expressed by epithelia) across organoid genotypes. **B)** Ligand expression (*UCell* scores) by fibroblasts in co-cultures across organoid genotypes (Games-Howell pairwise test, n.s not significant). **C)** Receptor expression (*UCell* scores) by epithelia in co-cultures across organoid genotypes (Games-Howell pairwise test with Holm-adjusted *p*-values). **D-E)** EMD-PCA of epithelial cells regulated by exogenous ligands. **F)** PCA from **D)** coloured by EMD scores for CLU and LRIG1. Boxplots show min/max and quartiles. Red dot marks the mean value.

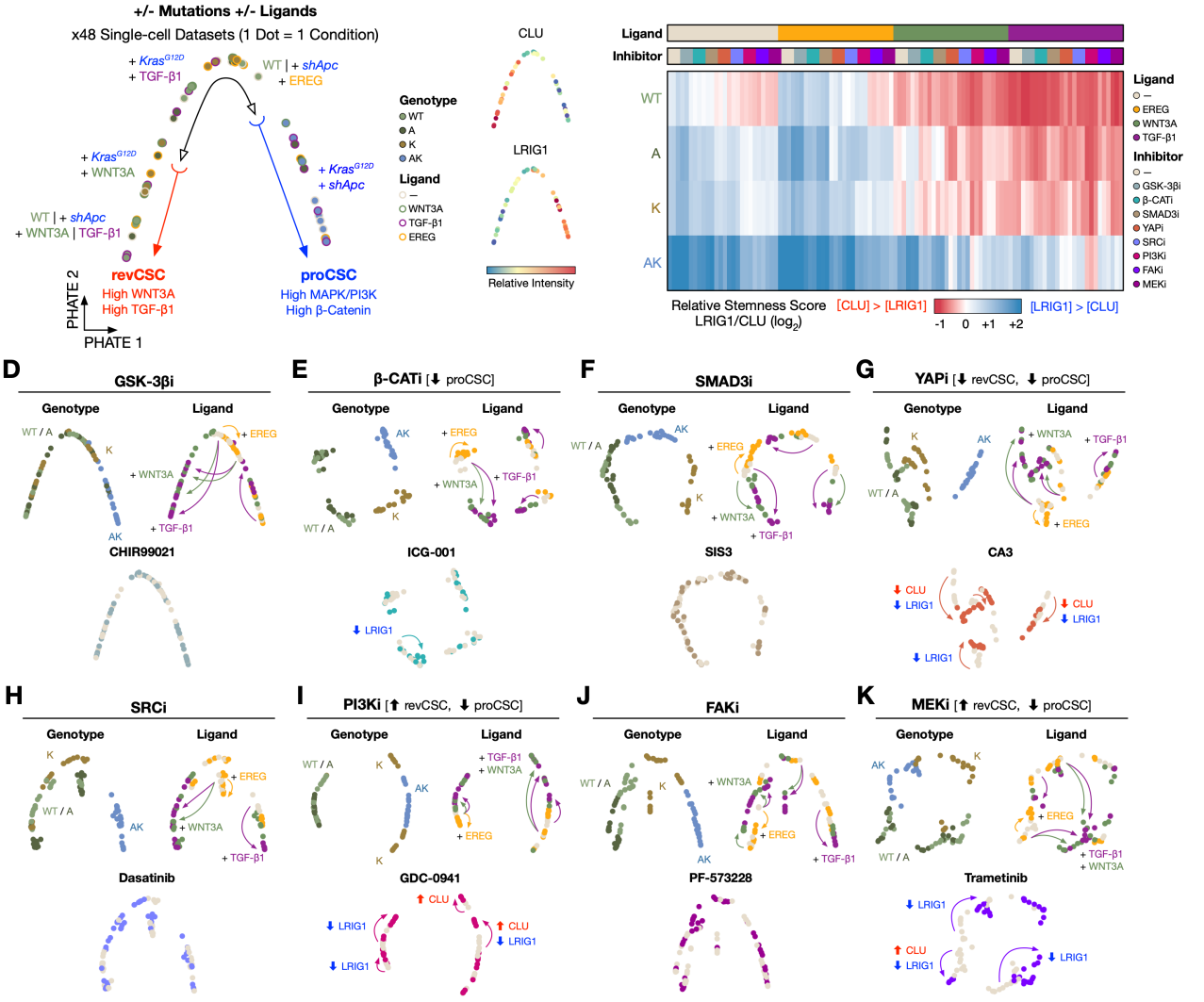


Figure S6. Signal Perturbation Analysis. Related to Figure 5. A) EMD-PHATE embedding of organoid cultures treated with ligands alone from the cue-signal-response array annotated with experimental metadata. One dot = one condition. **B)** PHATE embedding from **A)** coloured by EMD scores for CLU and LRIG1. **C)** Heatmap of relative stemness scores (log_2-transformed single-cell expression ratio between LRIG1 and CLU) of 432 organoid cultures from the cue-signal-response array. **D-K)** EMD-PHATE embeddings of organoid culture subsets from the cue-signal-response array focusing on each inhibitor.

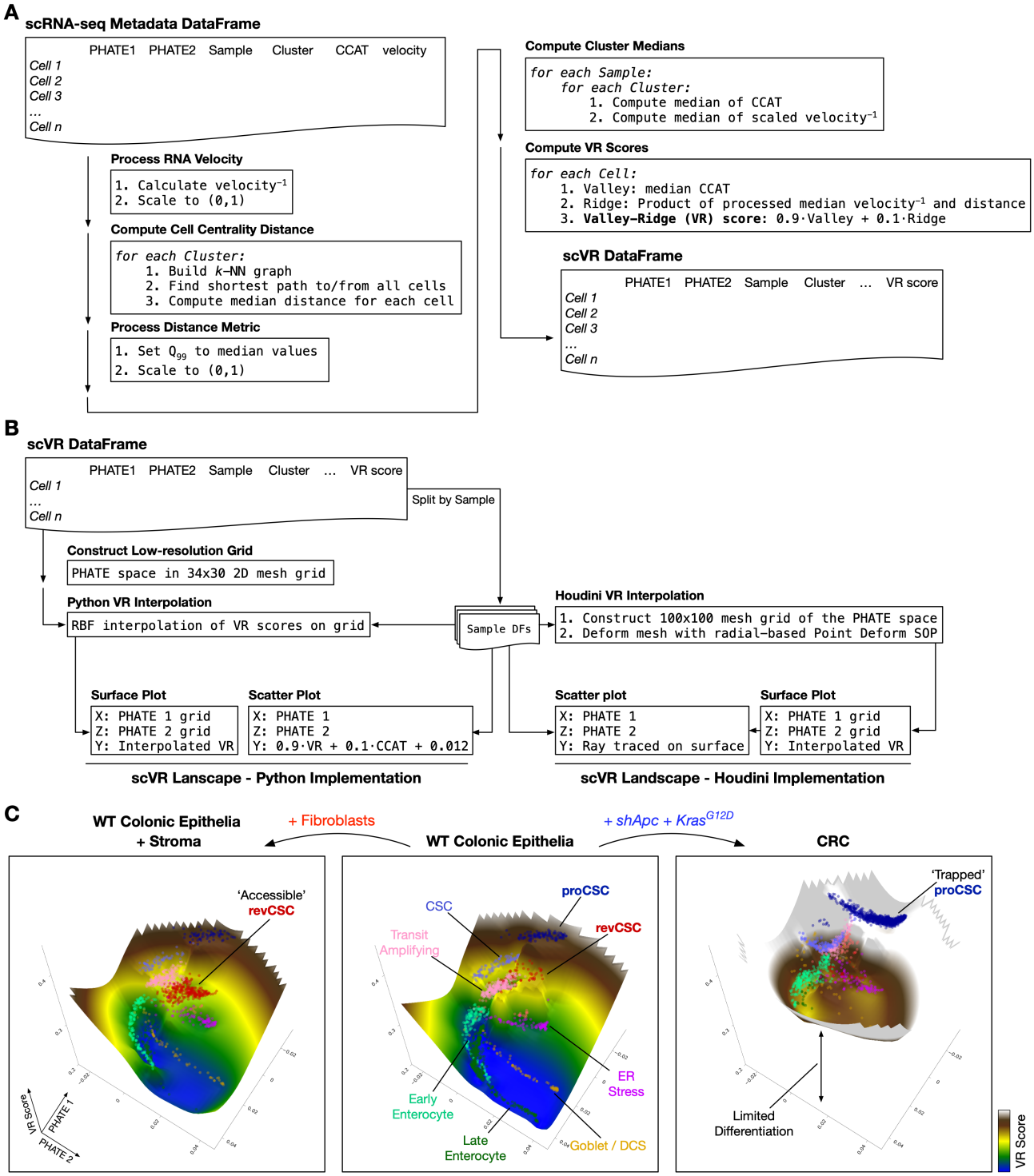


Figure S7. Generation of Waddington-like Landscapes from scRNA-seq Data. Related to Figure 6. A) Workflow for calculating VR scores from scRNA-seq data. **B)** Workflow for visualising PHATE and VR scores as 3D landscapes with either Python (left) or Houdini (right). RBF, radial basis function. **C)** PHATE and VR score landscapes visualised in Python.

Appendix E

Sufi & Qin *et al.*, 2021

Appendix F

Colophon

This Thesis has been written with LaTeXand its figures assembled using the FOSS vector graphics editor Inkscape (`inkscape.org`). Resources used for making figures include plots generated from code (R and Python), *de novo* drawn graphics, and graphics altered from the open source Bioicons resource (`bioicons.com`).

The Thesis is currently hosted in GitHub as a private repository. However, once the sections of work currently under revision at Cell are part of the public domain, I will make the repository public. Code availability covers all chapters and is distributed along multiple GitHub repositories.

Bibliography

[1] Ferran Cardoso Rodriguez, Xiao Qin, Jahangir Sufi, Petra Vlckova, Jeroen Claus, and Christopher J. Tape. A Single-cell Perturbation Landscape of Colonic Stem Cell Polarisation, February 2023. Pages: 2023.02.15.528008 Section: New Results.

[2] Jahangir Sufi, Xiao Qin, Ferran Cardoso Rodriguez, Yong Jia Bu, Petra Vlckova, María Ramos Zapatero, Mark Nitz, and Christopher J. Tape. Multiplexed single-cell analysis of organoid signaling networks. *Nature Protocols*, 16(10):4897–4918, October 2021. Number: 10 Publisher: Nature Publishing Group.

[3] María Ramos Zapatero, Alexander Tong, Jahangir Sufi, Petra Vlckova, Ferran Cardoso Rodriguez, Callum Nattress, Xiao Qin, Daniel Hochhauser, Smita Krishnaswamy, and Christopher J. Tape. Trellis Single-Cell Screening Reveals Stromal Regulation of Patient-Derived Organoid Drug Responses, January 2023. Pages: 2022.10.19.512668 Section: New Results.

[4] Xiao Qin, Jahangir Sufi, Petra Vlckova, Pelagia Kyriakidou, Sophie E. Acton, Vivian S. W. Li, Mark Nitz, and Christopher J. Tape. Cell-type-specific signaling networks in heterocellular organoids. *Nature Methods*, pages 1–8, February 2020.

[5] Massimo Andreatta and Santiago J. Carmona. UCell: Robust and scalable single-cell gene signature scoring. *Computational and Structural Biotechnology Journal*, 19:3796–3798, January 2021. Publisher: Elsevier.

[6] Ignasius Joanito, Pratyaksha Wirapati, Nancy Zhao, Zahid Nawaz, Grace Yeo, Fiona Lee, Christine L. P. Eng, Dominique Camat Macalinao, Merve Kahraman, Harini Srinivasan, Vairavan Lakshmanan, Sara Verbandt, Petros Tsantoulis, Nicole Gunn, Prasanna Nori Venkatesh, Zhong Wee Poh, Rahul Nahar, Hsueh Ling Janice Oh, Jia Min Loo, Shumei Chia, Lih Feng Cheow, Elsie Cheruba, Michael Thomas Wong, Lindsay Kua, Clarinda Chua, Andy Nguyen, Justin Golovan, Anna Gan, Wan-Jun Lim, Yu Amanda Guo, Choon Kong Yap, Brenda Tay, Yourae Hong, Dawn Qingqing Chong, Aik-Yong Chok, Woong-Yang Park, Shuting Han, Mei Huan Chang, Isaac Seow-En, Cherylin Fu, Ronnie Mathew, Ee-Lin Toh, Lewis Z. Hong, Anders Jacobsen Skanderup, Ramanuj DasGupta, Chin-Ann Johnny Ong, Kiat Hon Lim, Emile K. W. Tan, Si-Lin Koo, Wei Qiang Leow, Sabine Tejpar, Shyam Prabhakar, and Iain Beehuat Tan. Single-cell and bulk transcriptome sequencing identifies two epithelial tumor cell states and refines the consensus molecular classification of colorectal cancer. *Nature Genetics*, 54(7):963–975, July 2022. Number: 7 Publisher: Nature Publishing Group.

[7] Dénes Türei, Alberto Valdeolivas, Lejla Gul, Nicolàs Palacio-Escat, Michal Klein, Olga Ivanova, Márton Ölbei, Attila Gábor, Fabian Theis, Dezső Módos, Tamás Korcsmáros, and Julio Saez-Rodriguez. Integrated intra- and intercellular signaling knowledge for multicellular omics analysis. *Molecular Systems Biology*, 17(3):e9923, March 2021. Publisher: John Wiley & Sons, Ltd.

[8] Itay Tirosh, Benjamin Izar, Sanjay M. Prakadan, Marc H. Wadsworth, Daniel Treacy, John J. Trombetta, Asaf Rotem, Christopher Rodman, Christine Lian, George Murphy, Mohammad Fallahi-Sichani, Ken Dutton-Regester, Jia-Ren Lin, Ofir Cohen, Parin Shah, Diana Lu, Alex S. Genshaft, Travis K. Hughes, Carly G. K. Ziegler, Samuel W. Kazer, Aleth Gaillard, Kellie E. Kolb, Alexandra-Chloé Villani, Cory M. Johannessen, Aleksandr Y. Andreev, Eliezer M. Van Allen, Monica Bertagnolli, Peter K. Sorger, Ryan J. Sullivan, Keith T. Flaherty, Dennie T. Frederick, Judit Jané-Valbuena, Charles H. Yoon, Orit Rozenblatt-Rosen, Alex K. Shalek, Aviv Regev, and Levi A. Garraway.

Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-seq. *Science*, 352(6282):189–196, April 2016. Publisher: American Association for the Advancement of Science Section: Research Article.

[9] Evan Z. Macosko, Anindita Basu, Rahul Satija, James Nemesh, Karthik Shekhar, Melissa Goldman, Itay Tirosh, Allison R. Bialas, Nolan Kamitaki, Emily M. Martersteck, John J. Trombetta, David A. Weitz, Joshua R. Sanes, Alex K. Shalek, Aviv Regev, and Steven A. McCarroll. Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets. *Cell*, 161(5):1202–1214, May 2015. Publisher: Elsevier.

[10] Kathleen A. Cronin, Susan Scott, Albert U. Firth, Hyuna Sung, S. Jane Henley, Recinda L. Sherman, Rebecca L. Siegel, Robert N. Anderson, Betsy A. Kohler, Vicki B. Benard, Serban Negoita, Charles Wiggins, William G. Cance, and Ahmedin Jemal. Annual report to the nation on the status of cancer, part 1: National cancer statistics. *Cancer*, 128(24):4251–4284, 2022. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/cncr.34479.

[11] Eileen Morgan, Melina Arnold, A. Gini, V. Lorenzoni, C. J. Cabasag, Mathieu Laversanne, Jerome Vignat, Jacques Ferlay, Neil Murphy, and Freddie Bray. Global burden of colorectal cancer in 2020 and 2040: incidence and mortality estimates from GLOBOCAN. *Gut*, 72(2):338–344, February 2023. Publisher: BMJ Publishing Group Section: Colon.

[12] Eric R. Fearon and Bert Vogelstein. A genetic model for colorectal tumorigenesis. *Cell*, 61(5):759–767, June 1990.

[13] Amirsaeed Sabeti Aghabozorgi, Amirhossein Bahreyni, Atena Soleimani, Afsane Bahrami, Majid Khazaei, Gordon A. Ferns, Amir Avan, and Seyed Mahdi Hassanian. Role of adenomatous polyposis coli (APC) gene mutations in the pathogenesis of colorectal cancer; current status and perspectives. *Biochimie*, 157:64–71, February 2019.

[14] Tannaz Armaghany, Jon D. Wilson, Quyen Chu, and Glenn Mills. Genetic Alterations in Colorectal Cancer. *Gastrointestinal Cancer Research : GCR*, 5(1):19–27, 2012.

[15] Vijay G. Peddareddigari, Dingzhi Wang, and Raymond N. DuBois. The Tumor Microenvironment in Colorectal Carcinogenesis. *Cancer Microenvironment*, 3(1):149–166, March 2010.

[16] Vangelis Bonis, Carla Rossell, and Helmuth Gehart. The Intestinal Epithelium – Fluid Fate and Rigid Structure From Crypt Bottom to Villus Tip. *Frontiers in Cell and Developmental Biology*, 9, 2021.

[17] Pawel R. Kiela and Fayez K. Ghishan. Physiology of Intestinal Absorption and Secretion. *Best Practice & Research Clinical Gastroenterology*, 30(2):145–159, April 2016.

[18] Jarom Heijmans, Jooske F. van Lidth de Jeude, Bon-Kyoung Koo, Sanne L. Rosekrans, Mattheus C. B. Wielenga, Marc van de Wetering, Marc Ferrante, Amy S. Lee, Jos J. M. Onderwater, James C. Paton, Adrienne W. Paton, A. Mieke Mommaas, Liudmila L. Kodach, James C. Hardwick, Daniël W. Hommes, Hans Clevers, Vanesa Muncan, and Gijs R. van den Brink. ER Stress Causes Rapid Loss of Intestinal Epithelial Stemness through Activation of the Unfolded Protein Response. *Cell Reports*, 3(4):1128–1139, April 2013.

[19] Olivia I. Coleman and Dirk Haller. ER Stress and the UPR in Shaping Intestinal Tissue Homeostasis and Immunity. *Frontiers in Immunology*, 10, 2019.

[20] Joep Beumer and Hans Clevers. Cell fate specification and differentiation in the adult mammalian intestine. *Nature Reviews Molecular Cell Biology*, 22(1):39–53, January 2021. Number: 1 Publisher: Nature Publishing Group.

[21] Toshiro Sato, Daniel E. Stange, Marc Ferrante, Robert G. J. Vries, Johan H. van Es, Stieneke van den Brink, Winan J. van Houdt, Apollo Pronk,

Joost van Gorp, Peter D. Siersema, and Hans Clevers. Long-term Expansion of Epithelial Organoids From Human Colon, Adenoma, Adenocarcinoma, and Barrett's Epithelium. *Gastroenterology*, 141(5):1762–1772, November 2011. Publisher: Elsevier.

[22] Ayano Kondo and Klaus H. Kaestner. Emerging diverse roles of telocytes. *Development*, 146(14):dev175018, July 2019.

[23] Nobuo Sasaki, Norman Sachs, Kay Wiebrands, Saskia I. J. Ellenbroek, Arianna Fumagalli, Anna Lyubimova, Harry Begthel, Maaike van den Born, Johan H. van Es, Wouter R. Karthaus, Vivian S. W. Li, Carmen López-Iglesias, Peter J. Peters, Jacco van Rheenen, Alexander van Oudenaarden, and Hans Clevers. Reg4+ deep crypt secretory cells function as epithelial niche for Lgr5+ stem cells in colon. *Proceedings of the National Academy of Sciences*, 113(37):E5399–E5407, September 2016. Publisher: Proceedings of the National Academy of Sciences.

[24] Nathalie Sphyris, Michael C. Hodder, and Owen J. Sansom. Subversion of Niche-Signalling Pathways in Colorectal Cancer: What Makes and Breaks the Intestinal Stem Cell. *Cancers*, 13(5):1000, January 2021. Number: 5 Publisher: Multidisciplinary Digital Publishing Institute.

[25] Marc van de Wetering, Elena Sancho, Cornelis Verweij, Wim de Lau, Irma Oving, Adam Hurlstone, Karin van der Horn, Eduard Batlle, Damien Coudreuse, Anna-Pavlina Haramis, Menno Tjon-Pon-Fong, Petra Moerer, Maaike van den Born, Gwen Soete, Steven Pals, Martin Eilers, Rene Medema, and Hans Clevers. The β-Catenin/TCF-4 Complex Imposes a Crypt Progenitor Phenotype on Colorectal Cancer Cells. *Cell*, 111(2):241–250, October 2002.

[26] Maxim Norkin, Claudia Capdevila, Ruben I. Calderon, Tianhong Su, Maria Trifas, Paloma Ordóñez-Morán, and Kelley S. Yan. Single-Cell Studies of Intestinal Stem Cell Heterogeneity During Homeostasis and Regeneration. In Paloma Ordóñez-Morán, editor, *Intestinal Stem Cells: Methods and Protocols*,

Methods in Molecular Biology, pages 155–167. Springer US, New York, NY, 2020.

[27] Eric D. Bankaitis, Andrew Ha, Calvin J. Kuo, and Scott T. Magness. Reserve Stem Cells in Intestinal Homeostasis and Injury. *Gastroenterology*, 155(5):1348–1361, November 2018.

[28] Francisco M. Barriga, Elisa Montagni, Miyeko Mana, Maria Mendez-Lago, Xavier Hernando-Momblona, Marta Sevillano, Amy Guillaumet-Adkins, Gustavo Rodriguez-Esteban, Simon J. A. Buczacki, Marta Gut, Holger Heyn, Douglas J. Winton, Omer H. Yilmaz, Camille Stephan-Otto Attolini, Ivo Gut, and Eduard Batlle. Mex3a Marks a Slowly Dividing Subpopulation of Lgr5+ Intestinal Stem Cells. *Cell Stem Cell*, 20(6):801–816.e7, June 2017.

[29] Johannes Bues, Marjan Biočanin, Joern Pezoldt, Riccardo Dainese, Antonius Chrisnandy, Saba Rezakhani, Wouter Saelens, Vincent Gardeux, Revant Gupta, Rita Sarkis, Julie Russeil, Yvan Saeys, Esther Amstad, Manfred Claassen, Matthias P. Lutolf, and Bart Deplancke. Deterministic scRNA-seq captures variation in intestinal crypt and organoid composition. *Nature Methods*, 19(3):323–330, March 2022. Number: 3 Publisher: Nature Publishing Group.

[30] Arshad Ayyaz, Sandeep Kumar, Bruno Sangiorgi, Bibaswan Ghoshal, Jessica Gosio, Shaida Ouladan, Mardi Fink, Seda Barutcu, Daniel Trcka, Jess Shen, Kin Chan, Jeffrey L. Wrana, and Alex Gregorieff. Single-cell transcriptomes of the regenerating intestine reveal a revival stem cell. *Nature*, 569(7754):121–125, May 2019. Bandiera_abtest: a Cg_type: Nature Research Journals Number: 7754 Primary_atype: Research Publisher: Nature Publishing Group Subject_term: Intestinal stem cells;Morphogenesis Subject_term_id: intestinal-stem-cells;morphogenesis.

[31] Sumaiyah K. Rehman, Jennifer Haynes, Evelyne Collignon, Kevin R. Brown, Yadong Wang, Allison M. L. Nixon, Jeffrey P. Bruce, Jeffrey A. Wintersinger,

Arvind Singh Mer, Edwyn B. L. Lo, Cherry Leung, Evelyne Lima-Fernandes, Nicholas M. Pedley, Fraser Soares, Sophie McGibbon, Housheng Hansen He, Aaron Pollet, Trevor J. Pugh, Benjamin Haibe-Kains, Quaid Morris, Miguel Ramalho-Santos, Sidhartha Goyal, Jason Moffat, and Catherine A. O'Brien. Colorectal Cancer Cells Enter a Diapause-like DTP State to Survive Chemotherapy. *Cell*, 184(1):226–242.e21, January 2021.

[32] Adrián Álvarez Varela, Laura Novellasdemunt, Francisco M. Barriga, Xavier Hernando-Momblona, Adrià Cañellas-Socias, Sara Cano-Crespo, Marta Sevillano, Carme Cortina, Diana Stork, Clara Morral, Gemma Turon, Felipe Slebe, Laura Jiménez-Gracia, Ginevra Caratù, Peter Jung, Giorgio Stassi, Holger Heyn, Daniele V. F. Tauriello, Lidia Mateo, Sabine Tejpar, Elena Sancho, Camille Stephan-Otto Attolini, and Eduard Batlle. Mex3a marks drug-tolerant persister colorectal cancer cells that mediate relapse after chemotherapy. *Nature Cancer*, pages 1–19, June 2022. Publisher: Nature Publishing Group.

[33] Frances R. Balkwill, Melania Capasso, and Thorsten Hagemann. The tumor microenvironment at a glance. *Journal of Cell Science*, 125(23):5591–5596, December 2012.

[34] Alexandre Calon, Enza Lonardo, Antonio Berenguer-Llergo, Elisa Espinet, Xavier Hernando-Momblona, Mar Iglesias, Marta Sevillano, Sergio Palomo-Ponce, Daniele V. F. Tauriello, Daniel Byrom, Carme Cortina, Clara Morral, Carles Barceló, Sebastien Tosi, Antoni Riera, Camille Stephan-Otto Attolini, David Rossell, Elena Sancho, and Eduard Batlle. Stromal gene expression defines poor-prognosis subtypes in colorectal cancer. *Nature Genetics*, 47(4):320–329, April 2015.

[35] Claudio Isella, Andrea Terrasi, Sara Erika Bellomo, Consalvo Petti, Giovanni Galatola, Andrea Muratore, Alfredo Mellano, Rebecca Senetta, Adele Cassenti, Cristina Sonetto, Giorgio Inghirami, Livio Trusolino, Zsolt Fekete, Mark De Ridder, Paola Cassoni, Guy Storme, Andrea Bertotti, and Enzo

Medico. Stromal contribution to the colorectal cancer transcriptome. *Nature Genetics*, 47(4):312–319, April 2015.

[36] Christopher J. Tape. The Heterocellular Emergence of Colorectal Cancer. *Trends in Cancer*, 3(2):79–88, February 2017. Publisher: Elsevier.

[37] Salman M. Toor, Khaled Murshed, Mahmood Al-Dhaheri, Mahwish Khawar, Mohamed Abu Nada, and Eyad Elkord. Immune Checkpoints in Circulating and Tumor-Infiltrating CD4+ T Cell Subsets in Colorectal Cancer Patients. *Frontiers in Immunology*, 10, December 2019.

[38] George S. Karagiannis, Theofilos Poutahidis, Susan E. Erdman, Richard Kirsch, Robert H. Riddell, and Eleftherios P. Diamandis. Cancer-Associated Fibroblasts Drive the Progression of Metastasis through both Paracrine and Mechanical Pressure on Cancer Tissue. *Molecular Cancer Research*, 10(11):1403–1418, November 2012.

[39] Fernando O. Martinez and Siamon Gordon. The M1 and M2 paradigm of macrophage activation: time for reassessment. *F1000Prime Rep*, 6(13), March 2014.

[40] Meritxell Huch and Bon-Kyoung Koo. Modeling mouse and human development using organoid cultures. *Development*, 142(18):3113–3125, September 2015.

[41] Madeline A. Lancaster and Meritxell Huch. Disease modelling in human organoids. *Disease Models & Mechanisms*, 12(7), July 2019.

[42] Mohammad Almeqdadi, Miyeko D. Mana, Jatin Roper, and Ömer H. Yilmaz. Gut organoids: mini-tissues in culture to study intestinal physiology and disease. *American Journal of Physiology - Cell Physiology*, 317(3):C405–C419, September 2019.

[43] Toshiro Sato, Robert G. Vries, Hugo J. Snippert, Marc van de Wetering, Nick Barker, Daniel E. Stange, Johan H. van Es, Arie Abo, Pekka Kujala, Peter J.

Peters, and Hans Clevers. Single Lgr5 stem cells build crypt-villus structures in vitro without a mesenchymal niche. *Nature*, 459(7244):262–265, May 2009. Number: 7244 Publisher: Nature Publishing Group.

[44] Chang Su, Kelly A. Olsen, Catherine E. Bond, and Vicki L. J. Whitehall. The Efficacy of Using Patient-Derived Organoids to Predict Treatment Response in Colorectal Cancer. *Cancers*, 15(3):805, January 2023. Number: 3 Publisher: Multidisciplinary Digital Publishing Institute.

[45] Xiao Qin and Christopher J. Tape. Deciphering Organoids: High-Dimensional Analysis of Biomimetic Cultures. *Trends in Biotechnology*, December 2020.

[46] Lauren J. Tracey, Yeji An, and Monica J. Justice. CyTOF: An Emerging Technology for Single-Cell Proteomics in the Mouse. *Current Protocols*, 1(4):e118, 2021. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/cpz1.118.

[47] David Ochoa, Andrew F. Jarnuczak, Cristina Viéitez, Maja Gehre, Margaret Soucheray, André Mateus, Askar A. Kleefeldt, Anthony Hill, Luz Garcia-Alonso, Frank Stein, Nevan J. Krogan, Mikhail M. Savitski, Danielle L. Swaney, Juan A. Vizcaíno, Kyung-Min Noh, and Pedro Beltrao. The functional landscape of the human phosphoproteome. *Nature Biotechnology*, pages 1–9, December 2019.

[48] Sabine A. G. Cuijpers and Alfred C. O. Vertegaal. Guiding Mitotic Progression by Crosstalk between Post-translational Modifications. *Trends in Biochemical Sciences*, 43(4):251–268, April 2018.

[49] Suoqin Jin, Christian F. Guerrero-Juarez, Lihua Zhang, Ivan Chang, Raul Ramos, Chen-Hsiang Kuan, Peggy Myung, Maksim V. Plikus, and Qing Nie. Inference and analysis of cell-cell communication using CellChat. *Nature Communications*, 12(1):1088, February 2021. Number: 1 Publisher: Nature Publishing Group.

[50] Mirjana Efremova, Miquel Vento-Tormo, Sarah A. Teichmann, and Roser Vento-Tormo. CellPhoneDB: inferring cell–cell communication from com-

bined expression of multi-subunit ligand–receptor complexes. *Nature Protocols*, pages 1–23, February 2020. Publisher: Nature Publishing Group.

[51] Xinmin Li and Cun-Yu Wang. From bulk, single-cell to spatial RNA sequencing. *International Journal of Oral Science*, 13(1):1–6, November 2021. Number: 1 Publisher: Nature Publishing Group.

[52] Adam L. Haber, Moshe Biton, Noga Rogel, Rebecca H. Herbst, Karthik Shekhar, Christopher Smillie, Grace Burgin, Toni M. Delorey, Michael R. Howitt, Yarden Katz, Itay Tirosh, Semir Beyaz, Danielle Dionne, Mei Zhang, Raktima Raychowdhury, Wendy S. Garrett, Orit Rozenblatt-Rosen, Hai Ning Shi, Omer Yilmaz, Ramnik J. Xavier, and Aviv Regev. A single-cell survey of the small intestinal epithelium. *Nature*, 551(7680):333–339, November 2017.

[53] Jacob O. Kitzman. Haplotypes drop by drop. *Nature Biotechnology*, 34(3):296–298, March 2016. Number: 3 Publisher: Nature Publishing Group.

[54] Jiarui Ding, Xian Adiconis, Sean K. Simmons, Monika S. Kowalczyk, Cynthia C. Hession, Nemanja D. Marjanovic, Travis K. Hughes, Marc H. Wadsworth, Tyler Burks, Lan T. Nguyen, John Y. H. Kwon, Boaz Barak, William Ge, Amanda J. Kedaigle, Shaina Carroll, Shuqiang Li, Nir Hacohen, Orit Rozenblatt-Rosen, Alex K. Shalek, Alexandra-Chloé Villani, Aviv Regev, and Joshua Z. Levin. Systematic comparison of single-cell and single-nucleus RNA-sequencing methods. *Nature Biotechnology*, 38(6):737–746, June 2020. Number: 6 Publisher: Nature Publishing Group.

[55] Adam R. Abate, Chia-Hung Chen, Jeremy J. Agresti, and David A. Weitz. Beating Poisson encapsulation statistics using close-packed ordering. *Lab on a Chip*, 9(18):2628–2631, September 2009.

[56] Peng Qiu. Embracing the dropouts in single-cell RNA-seq analysis. *Nature Communications*, 11(1):1169, March 2020. Number: 1 Publisher: Nature Publishing Group.

[57] Kip D. Zimmerman, Mark A. Espeland, and Carl D. Langefeld. A practical solution to pseudoreplication bias in single-cell studies. *Nature Communications*, 12(1):738, February 2021. Number: 1 Publisher: Nature Publishing Group.

[58] Iain C. Clark, Kristina M. Fontanez, Robert H. Meltzer, Yi Xue, Corey Hayford, Aaron May-Zhang, Chris D'Amato, Ahmad Osman, Jesse Q. Zhang, Pabodha Hettige, Jacob S. A. Ishibashi, Cyrille L. Delley, Daniel W. Weisgerber, Joseph M. Replogle, Marco Jost, Kiet T. Phong, Vanessa E. Kennedy, Cheryl A. C. Peretz, Esther A. Kim, Siyou Song, William Karlon, Jonathan S. Weissman, Catherine C. Smith, Zev J. Gartner, and Adam R. Abate. Microfluidics-free single-cell genomics with templated emulsification. *Nature Biotechnology*, pages 1–10, March 2023. Publisher: Nature Publishing Group.

[59] Alexander B. Rosenberg, Charles M. Roco, Richard A. Muscat, Anna Kuchina, Paul Sample, Zizhen Yao, Lucas T. Graybuck, David J. Peeler, Sumit Mukherjee, Wei Chen, Suzie H. Pun, Drew L. Sellers, Bosiljka Tasic, and Georg Seelig. Single-cell profiling of the developing mouse brain and spinal cord with split-pool barcoding. *Science*, 360(6385):176–182, April 2018. Publisher: American Association for the Advancement of Science.

[60] KA Wetterstrand. DNA Sequencing Costs: Data from the NHGRI Genome Sequencing Program (GSP), September 2022.

[61] Erika Check Hayden. Genome sequencing: the third generation. *Nature*, February 2009. Publisher: Nature Publishing Group.

[62] John Eid, Adrian Fehr, Jeremy Gray, Khai Luong, John Lyle, Geoff Otto, Paul Peluso, David Rank, Primo Baybayan, Brad Bettman, Arkadiusz Bibillo, Keith Bjornson, Bidhan Chaudhuri, Frederick Christians, Ronald Cicero, Sonya Clark, Ravindra Dalal, Alex deWinter, John Dixon, Mathieu Foquet, Alfred Gaertner, Paul Hardenbol, Cheryl Heiner, Kevin Hester, David Holden, Gregory Kearns, Xiangxu Kong, Ronald Kuse, Yves Lacroix, Steven Lin, Paul

Lundquist, Congcong Ma, Patrick Marks, Mark Maxham, Devon Murphy, Insil Park, Thang Pham, Michael Phillips, Joy Roy, Robert Sebra, Gene Shen, Jon Sorenson, Austin Tomaney, Kevin Travers, Mark Trulson, John Vieceli, Jeffrey Wegener, Dawn Wu, Alicia Yang, Denis Zaccarin, Peter Zhao, Frank Zhong, Jonas Korlach, and Stephen Turner. Real-Time DNA Sequencing from Single Polymerase Molecules. *Science*, 323(5910):133–138, January 2009. Publisher: American Association for the Advancement of Science.

[63] David Deamer, Mark Akeson, and Daniel Branton. Three decades of nanopore sequencing. *Nature Biotechnology*, 34(5):518–524, May 2016. Number: 5 Publisher: Nature Publishing Group.

[64] Peng Ni, Neng Huang, Zhi Zhang, De-Peng Wang, Fan Liang, Yu Miao, Chuan-Le Xiao, Feng Luo, and Jianxin Wang. DeepSignal: detecting DNA methylation state from Nanopore sequencing reads using deep-learning. *Bioinformatics*, 35(22):4586–4595, November 2019.

[65] Laura Mincarelli, Ashleigh Lister, James Lipscombe, and Iain C. Macaulay. Defining Cell Identity with Single-Cell Omics. *PROTEOMICS*, 18(18):1700312, 2018. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/pmic.201700312.

[66] Sanja Vickovic, Gökcen Eraslan, Fredrik Salmén, Johanna Klughammer, Linnea Stenbeck, Denis Schapiro, Tarmo Äijö, Richard Bonneau, Ludvig Bergenstråhle, José Fernandéz Navarro, Joshua Gould, Gabriel K. Griffin, Åke Borg, Mostafa Ronaghi, Jonas Frisén, Joakim Lundeberg, Aviv Regev, and Patrik L. Ståhl. High-definition spatial transcriptomics for in situ tissue profiling. *Nature Methods*, 16(10):987–990, October 2019.

[67] Vivien Marx. Method of the Year: spatially resolved transcriptomics. *Nature Methods*, 18(1):9–14, January 2021. Number: 1 Publisher: Nature Publishing Group.

[68] Cameron G. Williams, Hyun Jae Lee, Takahiro Asatsuma, Roser Vento-Tormo, and Ashraful Haque. An introduction to spatial transcriptomics for biomedical research. *Genome Medicine*, 14(1):68, June 2022.

[69] Roman Hornung, Anne-Laure Boulesteix, and David Causeur. Combining location-and-scale batch effect adjustment with data cleaning by latent factor adjustment. *BMC Bioinformatics*, 17(1):27, January 2016.

[70] Andrew Butler, Paul Hoffman, Peter Smibert, Efthymia Papalexi, and Rahul Satija. Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nature Biotechnology*, 36(5):411–420, May 2018. Number: 5 Publisher: Nature Publishing Group.

[71] Massimo Andreatta and Santiago J Carmona. STACAS: Sub-Type Anchor Correction for Alignment in Seurat to integrate single-cell RNA-seq data. *Bioinformatics*, 37(6):882–884, March 2021.

[72] Huidong Chen, Caleb Lareau, Tommaso Andreani, Michael E. Vinyard, Sara P. Garcia, Kendell Clement, Miguel A. Andrade-Navarro, Jason D. Buenrostro, and Luca Pinello. Assessment of computational methods for the analysis of single-cell ATAC-seq data. *Genome Biology*, 20(1):241, November 2019.

[73] Kai Cao, Xiangqi Bai, Yiguang Hong, and Lin Wan. Unsupervised topological alignment for single-cell multi-omics integration. *Bioinformatics*, 36(Supplement_1):i48–i56, July 2020. Publisher: Oxford Academic.

[74] Zhi-Jie Cao and Ge Gao. Multi-omics single-cell data integration and regulatory inference with graph-linked embedding. *Nature Biotechnology*, 40(10):1458–1466, October 2022.

[75] Matthew Amodio and Smita Krishnaswamy. MAGAN: Aligning Biological Manifolds. *arXiv:1803.00385 [cs]*, February 2018. arXiv: 1803.00385.

[76] Mohammad Lotfollahi, Mohsen Naghipourfar, Malte D. Luecken, Matin Khajavi, Maren Büttner, Marco Wagenstetter, Žiga Avsec, Adam Gayoso,

Nir Yosef, Marta Interlandi, Sergei Rybakov, Alexander V. Misharin, and Fabian J. Theis. Mapping single-cell data to reference atlases by transfer learning. *Nature Biotechnology*, 40(1):121–130, January 2022. Number: 1 Publisher: Nature Publishing Group.

[77] Malte D Luecken and Fabian J Theis. Current best practices in single-cell RNA-seq analysis: a tutorial. *Molecular Systems Biology*, 15(6):e8746, June 2019.

[78] Lukas Heumos, Anna C. Schaar, Christopher Lance, Anastasia Litinetskaya, Felix Drost, Luke Zappia, Malte D. Lücken, Daniel C. Strobl, Juan Henao, Fabiola Curion, Herbert B. Schiller, and Fabian J. Theis. Best practices for single-cell analysis across modalities. *Nature Reviews Genetics*, pages 1–23, March 2023. Publisher: Nature Publishing Group.

[79] Alexander Dobin, Carrie A. Davis, Felix Schlesinger, Jorg Drenkow, Chris Zaleski, Sonali Jha, Philippe Batut, Mark Chaisson, and Thomas R. Gingeras. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics (Oxford, England)*, 29(1):15–21, January 2013.

[80] Nicolas L. Bray, Harold Pimentel, Páll Melsted, and Lior Pachter. Near-optimal probabilistic RNA-seq quantification. *Nature Biotechnology*, 34(5):525–527, May 2016. Number: 5 Publisher: Nature Publishing Group.

[81] Sean C. Bendall, Erin F. Simonds, Peng Qiu, El-ad D. Amir, Peter O. Krutzik, Rachel Finck, Robert V. Bruggner, Rachel Melamed, Angelica Trejo, Olga I. Ornatsky, Robert S. Balderas, Sylvia K. Plevritis, Karen Sachs, Dana Pe'er, Scott D. Tanner, and Garry P. Nolan. Single-cell mass cytometry of differential immune and drug responses across a human hematopoietic continuum. *Science (New York, N.Y.)*, 332(6030):687–696, May 2011.

[82] Sophia M. Guldberg, Trine Line Hauge Okholm, Elizabeth E. McCarthy, and Matthew H. Spitzer. Computational Methods for Single-Cell Pro-

teomics. *Annual Review of Biomedical Data Science*, 6(1):null, 2023. _eprint: https://doi.org/10.1146/annurev-biodatasci-020422-050255.

[83] Constantin Ahlmann-Eltze and Wolfgang Huber. Comparison of transformations for single-cell RNA-seq data. *Nature Methods*, pages 1–8, April 2023. Publisher: Nature Publishing Group.

[84] A. Sina Booeshaghi, Ingileif B. Hallgrímsdóttir, Ángel Gálvez-Merchán, and Lior Pachter. Depth normalization for single-cell genomics count data, May 2022. Pages: 2022.05.06.490859 Section: New Results.

[85] Christoph Hafemeister and Rahul Satija. Normalization and variance stabilization of single-cell RNA-seq data using regularized negative binomial regression. *Genome Biology*, 20(1):296, December 2019.

[86] Gregory K. Behbehani, Sean C. Bendall, Matthew R. Clutter, Wendy J. Fantl, and Garry P. Nolan. Single-cell mass cytometry adapted to measurements of the cell cycle. *Cytometry Part A*, 81A(7):552–566, 2012. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/cyto.a.22075.

[87] Karl Pearson. LIII. On lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 2(11):559–572, November 1901. Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/14786440109462720.

[88] Ian T. Jolliffe and Jorge Cadima. Principal component analysis: a review and recent developments. *Philosophical transactions. Series A, Mathematical, physical, and engineering sciences*, 374(2065):20150202, April 2016.

[89] Ronald R. Coifman and Mauro Maggioni. Diffusion wavelets. *Applied and Computational Harmonic Analysis*, 21(1):53–94, July 2006.

[90] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-SNE. *The Journal of Machine Learning Research*, 9(2579-2605):85, November 2008.

[91] Leland McInnes, John Healy, and James Melville. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. *arXiv:1802.03426 [cs, stat]*, September 2020. arXiv: 1802.03426.

[92] Kevin R. Moon, David van Dijk, Zheng Wang, Scott Gigante, Daniel B. Burkhardt, William S. Chen, Kristina Yim, Antonia van den Elzen, Matthew J. Hirn, Ronald R. Coifman, Natalia B. Ivanova, Guy Wolf, and Smita Krishnaswamy. Visualizing structure and transitions in high-dimensional biological data. *Nature Biotechnology*, 37(12):1482–1492, December 2019.

[93] Manik Kuchroo, Jessie Huang, Patrick Wong, Jean-Christophe Grenier, Dennis Shung, Alexander Tong, Carolina Lucas, Jon Klein, Daniel Burkhardt, Scott Gigante, Abhinav Godavarthi, Benjamin Israelow, Tianyang Mao, Ji Eun Oh, Julio Silva, Takehiro Takahashi, Camila D. Odio, Arnau Casanovas-Massana, John Fournier, Yale IMPACT Team, Shelli Farhadian, Charles S. Dela Cruz, Albert I. Ko, F. Perry Wilson, Julie Hussin, Guy Wolf, Akiko Iwasaki, and Smita Krishnaswamy. Multiscale PHATE Exploration of SARS-CoV-2 Data Reveals Multimodal Signatures of Disease. Technical report, November 2020. Company: Cold Spring Harbor Laboratory Distributor: Cold Spring Harbor Laboratory Label: Cold Spring Harbor Laboratory Section: New Results Type: article.

[94] Shaked Slovin, Annamaria Carissimo, Francesco Panariello, Antonio Grimaldi, Valentina Bouché, Gennaro Gambardella, and Davide Cacchiarelli. Single-Cell RNA Sequencing Analysis: A Step-by-Step Overview. In Ernesto Picardi, editor, *RNA Bioinformatics*, Methods in Molecular Biology, pages 343–365. Springer US, New York, NY, 2021.

[95] Vincent D. Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008(10):P10008, October 2008.

[96] V. A. Traag, L. Waltman, and N. J. van Eck. From Louvain to Leiden: guaranteeing well-connected communities. *Scientific Reports*, 9(1):5233, March 2019. Number: 1 Publisher: Nature Publishing Group.

[97] Kelly Street, Davide Risso, Russell B. Fletcher, Diya Das, John Ngai, Nir Yosef, Elizabeth Purdom, and Sandrine Dudoit. Slingshot: cell lineage and pseudotime inference for single-cell transcriptomics. *BMC Genomics*, 19(1):477, June 2018.

[98] F. Alexander Wolf, Fiona K. Hamey, Mireya Plass, Jordi Solana, Joakim S. Dahlin, Berthold Göttgens, Nikolaus Rajewsky, Lukas Simon, and Fabian J. Theis. PAGA: graph abstraction reconciles clustering with trajectory inference through a topology preserving map of single cells. *Genome Biology*, 20(1):59, March 2019.

[99] Gioele La Manno, Ruslan Soldatov, Amit Zeisel, Emelie Braun, Hannah Hochgerner, Viktor Petukhov, Katja Lidschreiber, Maria E. Kastriti, Peter Lönnerberg, Alessandro Furlan, Jean Fan, Lars E. Borm, Zehua Liu, David van Bruggen, Jimin Guo, Xiaoling He, Roger Barker, Erik Sundström, Gonçalo Castelo-Branco, Patrick Cramer, Igor Adameyko, Sten Linnarsson, and Peter V. Kharchenko. RNA velocity of single cells. *Nature*, 560(7719):494–498, August 2018. Number: 7719 Publisher: Nature Publishing Group.

[100] Volker Bergen, Marius Lange, Stefan Peidli, F. Alexander Wolf, and Fabian J. Theis. Generalizing RNA velocity to transient cell states through dynamical modeling. *Nature Biotechnology*, 38(12):1408–1414, December 2020. Number: 12 Publisher: Nature Publishing Group.

[101] Marius Lange, Volker Bergen, Michal Klein, Manu Setty, Bernhard Reuter, Mostafa Bakhti, Heiko Lickert, Meshal Ansari, Janine Schniering, Herbert B. Schiller, Dana Pe'er, and Fabian J. Theis. CellRank for directed single-cell fate mapping. *Nature Methods*, 19(2):159–170, February 2022. Number: 2 Publisher: Nature Publishing Group.

[102] Aaron T. L. Lun, Arianne C. Richard, and John C. Marioni. Testing for differential abundance in mass cytometry data. *Nature Methods*, 14(7):707–709, July 2017. Bandiera_abtest: a Cg_type: Nature Research Journals Number: 7 Primary_atype: Research Publisher: Nature Publishing Group Subject_term: Cytological techniques;Protein analysis;Statistical methods Subject_term_id: cytological-techniques;protein-analysis;statistical-methods.

[103] Emma Dann, Neil C. Henderson, Sarah A. Teichmann, Michael D. Morgan, and John C. Marioni. Differential abundance testing on single-cell data using k-nearest neighbor graphs. *Nature Biotechnology*, 40(2):245–253, February 2022. Number: 2 Publisher: Nature Publishing Group.

[104] Daniel B. Burkhardt, Jay S. Stanley, Alexander Tong, Ana Luisa Perdigoto, Scott A. Gigante, Kevan C. Herold, Guy Wolf, Antonio J. Giraldez, David van Dijk, and Smita Krishnaswamy. Quantifying the effect of experimental perturbations at single-cell resolution. *Nature Biotechnology*, 39(5):619–629, May 2021. Number: 5 Publisher: Nature Publishing Group.

[105] Alexander Tong, Jessie Huang, Guy Wolf, David van Dijk, and Smita Krishnaswamy. TrajectoryNet: A Dynamic Optimal Transport Network for Modeling Cellular Dynamics, July 2020. arXiv:2002.04461 [cs, q-bio, stat].

[106] Mohammad Lotfollahi, F. Alexander Wolf, and Fabian J. Theis. scGen predicts single-cell perturbation responses. *Nature Methods*, 16(8):715–721, August 2019. Number: 8 Publisher: Nature Publishing Group.

[107] Bo Yuan, Ciyue Shen, Augustin Luna, Anil Korkut, Debora S. Marks, John Ingraham, and Chris Sander. CellBox: Interpretable Machine Learning for Perturbation Biology with Application to the Design of Cancer Combination Therapy. *Cell Systems*, 12(2):128–140.e4, February 2021.

[108] Mohammad Lotfollahi, Anna Klimovskaia Susmelj, Carlo De Donno, Yuge Ji, Ignacio L. Ibarra, F. Alexander Wolf, Nafissa Yakubova, Fabian J. Theis, and

David Lopez-Paz. Learning interpretable cellular responses to complex perturbations in high-throughput screens, May 2021. Pages: 2021.04.14.439903 Section: New Results.

[109] Andrew E. Teschendorff, Peter Sollich, and Reimer Kuehn. Signalling entropy: A novel network-theoretical framework for systems analysis and interpretation of functional omic data. *Methods*, 67(3):282–293, June 2014.

[110] Andrew E. Teschendorff and Tariq Enver. Single-cell entropy for accurate estimation of differentiation potency from a cell's transcriptome. *Nature Communications*, 8(1):15599, June 2017. Number: 1 Publisher: Nature Publishing Group.

[111] Gunsagar S. Gulati, Shaheen S. Sikandar, Daniel J. Wesche, Anoop Manjunath, Anjan Bharadwaj, Mark J. Berger, Francisco Ilagan, Angera H. Kuo, Robert W. Hsieh, Shang Cai, Maider Zabala, Ferenc A. Scheeren, Neethan A. Lobo, Dalong Qian, Feiqiao B. Yu, Frederick M. Dirbas, Michael F. Clarke, and Aaron M. Newman. Single-cell transcriptional diversity is a hallmark of developmental potential. *Science*, 367(6476):405–411, January 2020. Publisher: American Association for the Advancement of Science.

[112] Daniela Senra, Nara Guisoni, and Luis Diambra. ORIGINS: A protein network-based approach to quantify cell pluripotency from scRNA-seq data. *MethodsX*, 9:101778, January 2022.

[113] Mark D. Robinson, Davis J. McCarthy, and Gordon K. Smyth. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, 26(1):139–140, January 2010.

[114] Greg Finak, Andrew McDavid, Masanao Yajima, Jingyuan Deng, Vivian Gersuk, Alex K. Shalek, Chloe K. Slichter, Hannah W. Miller, M. Juliana McElrath, Martin Prlic, Peter S. Linsley, and Raphael Gottardo. MAST: a flexible statistical framework for assessing transcriptional changes and

characterizing heterogeneity in single-cell RNA sequencing data. *Genome Biology*, 16(1):278, December 2015.

[115] Alsu Missarova, Emma Dann, Leah Rosen, Rahul Satija, and John Marioni. Sensitive cluster-free differential expression testing, March 2023. Pages: 2023.03.08.531744 Section: New Results.

[116] Michael Ashburner, Catherine A. Ball, Judith A. Blake, David Botstein, Heather Butler, J. Michael Cherry, Allan P. Davis, Kara Dolinski, Selina S. Dwight, Janan T. Eppig, Midori A. Harris, David P. Hill, Laurie Issel-Tarver, Andrew Kasarskis, Suzanna Lewis, John C. Matese, Joel E. Richardson, Martin Ringwald, Gerald M. Rubin, and Gavin Sherlock. Gene Ontology: tool for the unification of biology. *Nature Genetics*, 25(1):25–29, May 2000. Number: 1 Publisher: Nature Publishing Group.

[117] Minoru Kanehisa, Miho Furumichi, Mao Tanabe, Yoko Sato, and Kanae Morishima. KEGG: new perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Research*, 45(D1):D353–D361, January 2017.

[118] Marc Gillespie, Bijay Jassal, Ralf Stephan, Marija Milacic, Karen Rothfels, Andrea Senff-Ribeiro, Johannes Griss, Cristoffer Sevilla, Lisa Matthews, Chuqiao Gong, Chuan Deng, Thawfeek Varusai, Eliot Ragueneau, Yusra Haider, Bruce May, Veronica Shamovsky, Joel Weiser, Timothy Brunson, Nasim Sanati, Liam Beckman, Xiang Shao, Antonio Fabregat, Konstantinos Sidiropoulos, Julieth Murillo, Guilherme Viteri, Justin Cook, Solomon Shorser, Gary Bader, Emek Demir, Chris Sander, Robin Haw, Guanming Wu, Lincoln Stein, Henning Hermjakob, and Peter D'Eustachio. The reactome pathway knowledgebase 2022. *Nucleic Acids Research*, 50(D1):D687–D692, January 2022.

[119] Aravind Subramanian, Pablo Tamayo, Vamsi K. Mootha, Sayan Mukherjee, Benjamin L. Ebert, Michael A. Gillette, Amanda Paulovich, Scott L. Pomeroy, Todd R. Golub, Eric S. Lander, and Jill P. Mesirov. Gene set enrichment anal-

ysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences*, 102(43):15545–15550, October 2005. Publisher: Proceedings of the National Academy of Sciences.

[120] Sara Aibar, Carmen Bravo González-Blas, Thomas Moerman, Vân Anh Huynh-Thu, Hana Imrichova, Gert Hulselmans, Florian Rambow, Jean-Christophe Marine, Pierre Geurts, Jan Aerts, Joost van den Oord, Zeynep Kalender Atak, Jasper Wouters, and Stein Aerts. SCENIC: single-cell regulatory network inference and clustering. *Nature Methods*, 14(11):1083–1086, November 2017. Number: 11 Publisher: Nature Publishing Group.

[121] Erick Armingol, Adam Officer, Olivier Harismendy, and Nathan E. Lewis. Deciphering cell–cell interactions and communication from gene expression. *Nature Reviews Genetics*, pages 1–18, November 2020. Publisher: Nature Publishing Group.

[122] Yuxuan Hu, Tao Peng, Lin Gao, and Kai Tan. CytoTalk: De novo construction of signal transduction networks using single-cell transcriptomic data. *Science Advances*, 7(16):eabf1356, April 2021. Publisher: American Association for the Advancement of Science.

[123] Kazumasa Kanemaru, James Cranley, Daniele Muraro, Antonio M. A. Miranda, Jan Patrick Pett, Monika Litvinukova, Natsuhiko Kumasaka, Siew Yen Ho, Krzysztof Polanski, Laura Richardson, Lukas Mach, Monika Dabrowska, Nathan Richoz, Sam N. Barnett, Shani Perera, Anna Wilbrey-Clark, Carlos Talavera-López, Ilaria Mulas, Krishnaa T. Mahbubani, Liam Bolt, Lira Mamanova, Liz Tuck, Lu Wang, Margaret M. Huang, Martin Prete, Sophie Pritchard, John Dark, Kourosh Saeb-Parsy, Minal Patel, Menna R. Clatworthy, Norbert Hübner, Rasheda A. Chowdhury, Michela Noseda, and Sarah A. Teichmann. Spatially resolved multiomics of human cardiac niches, February 2023. Pages: 2023.01.30.526202 Section: New Results.

[124] Daniel Dimitrov, Dénes Türei, Martin Garrido-Rodriguez, Paul L. Burmedi, James S. Nagai, Charlotte Boys, Ricardo O. Ramirez Flores, Hyojin Kim, Bence Szalai, Ivan G. Costa, Alberto Valdeolivas, Aurélien Dugourd, and Julio Saez-Rodriguez. Comparison of methods and resources for cell-cell communication inference from single-cell RNA-Seq data. *Nature Communications*, 13(1):3224, June 2022. Number: 1 Publisher: Nature Publishing Group.

[125] Hratch M. Baghdassarian, Daniel Dimitrov, Erick Armingol, Julio Saez-Rodriguez, and Nathan E. Lewis. Combining LIANA and Tensor-cell2cell to decipher cell-cell communication across multiple samples, April 2023. Pages: 2023.04.28.538731 Section: New Results.

[126] Yael Baran, Akhiad Bercovich, Arnau Sebe-Pedros, Yaniv Lubling, Amir Giladi, Elad Chomsky, Zohar Meir, Michael Hoichman, Aviezer Lifshitz, and Amos Tanay. MetaCell: analysis of single-cell RNA-seq data using K-nn graph partitions. *Genome Biology*, 20(1):206, October 2019.

[127] Yuge Wang and Hongyu Zhao. Non-linear archetypal analysis of single-cell RNA-seq data by deep autoencoders. *PLOS Computational Biology*, 18(4):e1010025, 2022. Publisher: Public Library of Science.

[128] Sitara Persad, Zi-Ning Choo, Christine Dien, Noor Sohail, Ignas Masilionis, Ronan Chaligné, Tal Nawy, Chrysothemis C. Brown, Roshan Sharma, Itsik Pe'er, Manu Setty, and Dana Pe'er. SEACells infers transcriptional and epigenomic cellular states from single-cell genomics data. *Nature Biotechnology*, pages 1–12, March 2023. Publisher: Nature Publishing Group.

[129] C.H. Waddington. *Waddington The Strategy Of Genes 1957*. 1957.

[130] Weiyan Chen, Samuel J. Morabito, Kai Kessenbrock, Tariq Enver, Kerstin B. Meyer, and Andrew E. Teschendorff. Single-cell landscape in mammary epithelium reveals bipotent-like cells associated with breast cancer risk and

outcome. *Communications Biology*, 2(1):1–13, August 2019. Number: 1 Publisher: Nature Publishing Group.

[131] Chen Chen, Dabao Zhang, Tony R. Hazbun, and Min Zhang. Inferring Gene Regulatory Networks from a Population of Yeast Segregants. *Scientific Reports*, 9(1):1197, February 2019. Number: 1 Publisher: Nature Publishing Group.

[132] M Lefebvre, A Gaignard, M Folschette, J Bourdon, and C Guziolowski. Large-scale regulatory and signaling network assembly through linked open data. *Database*, 2021:baaa113, September 2021.

[133] Dirk Merkel. Docker: lightweight Linux containers for consistent development and deployment. *Linux Journal*, 2014(239):2:2, March 2014.

[134] Josef Spidlen, Wayne Moore, David Parks, Michael Goldberg, Chris Bray, Pierre Bierre, Peter Gorombey, Bill Hyun, Mark Hubbard, Simon Lange, Ray Lefebvre, Robert Leif, David Novo, Leo Ostruszka, Adam Treister, James Wood, Robert F. Murphy, Mario Roederer, Damir Sudar, Robert Zigon, and Ryan R. Brinkman. Data File Standard for Flow Cytometry, Version FCS 3.1. *Cytometry. Part A : the journal of the International Society for Analytical Cytology*, 77(1):97–100, January 2010.

[135] Eugene Yurtsev. eyurtsev/fcsparser, December 2020. original-date: 2015-09-13T22:09:34Z.

[136] ZELLMECHANIK-DRESDEN/fcswrite, January 2021. original-date: 2016-12-11T10:27:27Z.

[137] B. Ellis, Perry Haal, Florian Hahne, Nolwenn Le Meur, Nishant Gopalakrishnan, Josef Spidlen, Mike Jiang, Greg Finak, and Samuel Granjeaud. flowCore: flowCore: Basic structures for flow cytometry data, 2021.

[138] David van Dijk, Roshan Sharma, Juozas Nainys, Kristina Yim, Pooja Kathail, Ambrose J. Carr, Cassandra Burdziak, Kevin R. Moon, Christine L. Chaffer,

Diwakar Pattabiraman, Brian Bierie, Linas Mazutis, Guy Wolf, Smita Krishnaswamy, and Dana Pe'er. Recovering Gene Interactions from Single-Cell Data Using Data Diffusion. *Cell*, 174(3):716–729.e27, July 2018.

[139] KrishnaswamyLab/scprep, January 2021. original-date: 2017-08-08T17:41:39Z.

[140] rstudio/shiny, January 2021. original-date: 2012-06-20T18:45:11Z.

[141] Hadley Wickham. *ggplot2: Elegant Graphics for Data Analysis*. Use R! Springer-Verlag, New York, 2009.

[142] Zuguang Gu. ComplexHeatmap: Make Complex Heatmaps, 2021.

[143] Sébastien Lê, Julie Josse, and François Husson. FactoMineR: An R Package for Multivariate Analysis. *Journal of Statistical Software*, 25(1):1–18, March 2008. Number: 1.

[144] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12(85):2825–2830, 2011.

[145] 10X Genomics. What is Cell Ranger? -Software -Single Cell Gene Expression -Official 10x Genomics Support.

[146] Illumina. bcl2fastq Conversion Software.

[147] Yuhan Hao, Stephanie Hao, Erica Andersen-Nissen, William M. Mauck, Shiwei Zheng, Andrew Butler, Maddie J. Lee, Aaron J. Wilk, Charlotte Darby, Michael Zager, Paul Hoffman, Marlon Stoeckius, Efthymia Papalexi, Eleni P. Mimitou, Jaison Jain, Avi Srivastava, Tim Stuart, Lamar M. Fleming, Bertrand Yeung, Angela J. Rogers, Juliana M. McElrath, Catherine A. Blish, Raphael

Gottardo, Peter Smibert, and Rahul Satija. Integrated analysis of multimodal single-cell data. *Cell*, 184(13):3573–3587.e29, June 2021.

[148] Pierre-Luc Germain, Aaron Lun, Carlos Garcia Meixide, Will Macnair, and Mark D. Robinson. Doublet identification in single-cell sequencing data using *scDblFinder*. Technical Report 10:979, F1000Research, May 2022. Type: article.

[149] Tim Stuart and Rahul Satija. Integrative single-cell analysis. *Nature Reviews Genetics*, 20(5):257–272, May 2019.

[150] Romain Lopez, Jeffrey Regier, Michael B. Cole, Michael I. Jordan, and Nir Yosef. Deep generative modeling for single-cell transcriptomics. *Nature Methods*, 15(12):1053–1058, December 2018. Number: 12 Publisher: Nature Publishing Group.

[151] Yuyao Song, Zhichao Miao, Alvis Brazma, and Irene Papatheodorou. Benchmarking strategies for cross-species integration of single-cell RNA sequencing data, April 2023. Pages: 2022.09.27.509674 Section: New Results.

[152] Adam Gayoso, Romain Lopez, Galen Xing, Pierre Boyeau, Valeh Valiollah Pour Amiri, Justin Hong, Katherine Wu, Michael Jayasuriya, Edouard Mehlman, Maxime Langevin, Yining Liu, Jules Samaran, Gabriel Misrachi, Achille Nazaret, Oscar Clivio, Chenling Xu, Tal Ashuach, Mariano Gabitto, Mohammad Lotfollahi, Valentine Svensson, Eduardo da Veiga Beltrame, Vitalii Kleshchevnikov, Carlos Talavera-López, Lior Pachter, Fabian J. Theis, Aaron Streets, Michael I. Jordan, Jeffrey Regier, and Nir Yosef. A Python library for probabilistic analysis of single-cell omics data. *Nature Biotechnology*, 40(2):163–166, February 2022. Number: 2 Publisher: Nature Publishing Group.

[153] Stefan Peidli. mousipy, June 2023. original-date: 2022-01-10T16:08:43Z.

[154] Ferran Cardoso, Xiao Qin, and Christopher Tape. TAPE-Lab/CyGNAL: Release of v0.2, March 2021.

[155] Luke Zappia and Alicia Oshlack. Clustering trees: a visualization for evaluating clusterings at multiple resolutions. *GigaScience*, 7(7), July 2018.

[156] Zuguang Gu, Roland Eils, and Matthias Schlesner. Complex heatmaps reveal patterns and correlations in multidimensional genomic data. *Bioinformatics*, 32(18):2847–2849, September 2016.

[157] Ester Gil Vasquez, Nadia Nasreddin, Gabriel N. Valbuena, Eoghan J. Mulholland, Hayley L. Belnoue-Davis, Holly Eggington, Ryan O. Schenck, Valérie M. Wouters, Pratyaksha Wirapati, Kathryn Gilroy, Tamsin R. M. Lannagan, Dustin J. Flanagan, Arafath K. Najumudeen, Sulochana Omwenga, Amy M. B. McCorry, Alistair Easton, Viktor H. Koelzer, James E. East, Dion Morton, Livio Trusolino, Timothy Maughan, Andrew D. Campbell, Maurice B. Loughrey, Philip D. Dunne, Petros Tsantoulis, David J. Huels, Sabine Tejpar, Owen Sansom, and Simon J. Leedham. Molecular phenotyping of colorectal neoplasia shows dynamic and adaptive cancer stem cell population admixture, June 2022. Pages: 2022.06.11.495729 Section: New Results.

[158] Roxana C. Mustata, Gabriela Vasile, Valeria Fernandez-Vallone, Sandra Strollo, Anne Lefort, Frédérick Libert, Daniel Monteyne, David Pérez-Morga, Gilbert Vassart, and Marie-Isabelle Garcia. Identification of Lgr5-independent spheroid-generating progenitors of the mouse fetal intestinal epithelium. *Cell Reports*, 5(2):421–432, October 2013.

[159] Shiro Yui, Luca Azzolin, Martti Maimets, Marianne Terndrup Pedersen, Robert P. Fordham, Stine L. Hansen, Hjalte L. Larsen, Jordi Guiu, Mariana R. P. Alves, Carsten F. Rundsten, Jens V. Johansen, Yuan Li, Chris D. Madsen, Tetsuya Nakamura, Mamoru Watanabe, Ole H. Nielsen, Pawel J. Schweiger, Stefano Piccolo, and Kim B. Jensen. YAP/TAZ-Dependent Reprogramming of Colonic Epithelium Links ECM Remodeling to Tissue Regeneration. *Cell Stem Cell*, 22(1):35–49.e7, January 2018.

[160] Javier Muñoz, Daniel E Stange, Arnout G Schepers, Marc van de Wetering, Bon-Kyoung Koo, Shalev Itzkovitz, Richard Volckmann, Kevin S Kung, Jan Koster, Sorina Radulescu, Kevin Myant, Rogier Versteeg, Owen J Sansom, Johan H van Es, Nick Barker, Alexander van Oudenaarden, Shabaz Mohammed, Albert JR Heck, and Hans Clevers. The Lgr5 intestinal stem cell signature: robust expression of proposed quiescent '+4' cell markers. *The EMBO Journal*, 31(14):3079–3091, July 2012. Publisher: John Wiley & Sons, Ltd.

[161] Huipeng Li, Elise T. Courtois, Debarka Sengupta, Yuliana Tan, Kok Hao Chen, Jolene Jie Lin Goh, Say Li Kong, Clarinda Chua, Lim Kiat Hon, Wah Siew Tan, Mark Wong, Paul Jongjoon Choi, Lawrence J. K. Wee, Axel M. Hillmer, Iain Beehuat Tan, Paul Robson, and Shyam Prabhakar. Reference component analysis of single-cell transcriptomes elucidates cellular heterogeneity in human colorectal tumors. *Nature Genetics*, 49(5):708–718, May 2017. Number: 5 Publisher: Nature Publishing Group.

[162] Teng Han, Sukanya Goswami, Yang Hu, Fanying Tang, Maria Paz Zafra, Charles Murphy, Zhen Cao, John T. Poirier, Ekta Khurana, Olivier Elemento, Jaclyn F. Hechtman, Karuna Ganesh, Rona Yaeger, and Lukas E. Dow. Lineage Reversion Drives WNT Independence in Intestinal Cancer. *Cancer Discovery*, 10(10):1590–1609, October 2020.

[163] Anna Merlos-Suárez, Francisco M. Barriga, Peter Jung, Mar Iglesias, María Virtudes Céspedes, David Rossell, Marta Sevillano, Xavier Hernando-Momblona, Victoria da Silva-Diz, Purificación Muñoz, Hans Clevers, Elena Sancho, Ramón Mangues, and Eduard Batlle. The Intestinal Stem Cell Signature Identifies Colorectal Cancer Stem Cells and Predicts Disease Relapse. *Cell Stem Cell*, 8(5):511–524, May 2011.

[164] Piero Dalerba, Tomer Kalisky, Debashis Sahoo, Pradeep S. Rajendran, Michael E. Rothenberg, Anne A. Leyrat, Sopheak Sim, Jennifer Okamoto, Darius M. Johnston, Dalong Qian, Maider Zabala, Janet Bueno, Norma F. Neff, Jianbin Wang, Andrew A. Shelton, Brendan Visser, Shigeo Hisamori,

Yohei Shimono, Marc van de Wetering, Hans Clevers, Michael F. Clarke, and Stephen R. Quake. Single-cell dissection of transcriptional heterogeneity in human colon tumors. *Nature Biotechnology*, 29(12):1120–1127, December 2011. Number: 12 Publisher: Nature Publishing Group.

[165] Karin Pelka, Matan Hofree, Jonathan H. Chen, Siranush Sarkizova, Joshua D. Pirl, Vjola Jorgji, Alborz Bejnood, Danielle Dionne, William H. Ge, Katherine H. Xu, Sherry X. Chao, Daniel R. Zollinger, David J. Lieb, Jason W. Reeves, Christopher A. Fuhrman, Margaret L. Hoang, Toni Delorey, Lan T. Nguyen, Julia Waldman, Max Klapholz, Isaac Wakiro, Ofir Cohen, Julian Albers, Christopher S. Smillie, Michael S. Cuoco, Jingyi Wu, Mei-Ju Su, Jason Yeung, Brinda Vijaykumar, Angela M. Magnuson, Natasha Asinovski, Tabea Moll, Max N. Goder-Reiser, Anise S. Applebaum, Lauren K. Brais, Laura K. DelloStritto, Sarah L. Denning, Susannah T. Phillips, Emma K. Hill, Julia K. Meehan, Dennie T. Frederick, Tatyana Sharova, Abhay Kanodia, Ellen Z. Todres, Judit Jané-Valbuena, Moshe Biton, Benjamin Izar, Conner D. Lambden, Thomas E. Clancy, Ronald Bleday, Nelya Melnitchouk, Jennifer Irani, Hiroko Kunitake, David L. Berger, Amitabh Srivastava, Jason L. Hornick, Shuji Ogino, Asaf Rotem, Sébastien Vigneau, Bruce E. Johnson, Ryan B. Corcoran, Arlene H. Sharpe, Vijay K. Kuchroo, Kimmie Ng, Marios Giannakis, Linda T. Nieman, Genevieve M. Boland, Andrew J. Aguirre, Ana C. Anderson, Orit Rozenblatt-Rosen, Aviv Regev, and Nir Hacohen. Spatially organized multicellular immune hubs in human colorectal cancer. *Cell*, 184(18):4734–4752.e20, September 2021.

[166] Alex Gregorieff, Yu Liu, Mohammad R. Inanlou, Yuliya Khomchuk, and Jeffrey L. Wrana. Yap-dependent reprogramming of Lgr5+ stem cells drives intestinal regeneration and cancer. *Nature*, 526(7575):715–718, October 2015. Number: 7575 Publisher: Nature Publishing Group.

[167] Adrià Cañellas-Socias, Carme Cortina, Xavier Hernando-Momblona, Sergio Palomo-Ponce, Eoghan J. Mulholland, Gemma Turon, Lidia Mateo,

Sefora Conti, Olga Roman, Marta Sevillano, Felipe Slebe, Diana Stork, Adrià Caballé-Mestres, Antonio Berenguer-Llergo, Adrián Álvarez Varela, Nicola Fenderico, Laura Novellasdemunt, Laura Jiménez-Gracia, Tamara Sipka, Lidia Bardia, Patricia Lorden, Julien Colombelli, Holger Heyn, Xavier Trepat, Sabine Tejpar, Elena Sancho, Daniele V. F. Tauriello, Simon Leedham, Camille Stephan-Otto Attolini, and Eduard Batlle. Metastatic recurrence in colorectal cancer arises from residual EMP1+ cells. *Nature*, pages 1–11, November 2022. Publisher: Nature Publishing Group.

[168] Yumeng Wang, Xiaoyan Xu, Dejan Maglic, Michael T. Dill, Kamalika Mojumdar, Patrick Kwok-Shing Ng, Kang Jin Jeong, Yiu Huen Tsang, Daniela Moreno, Venkata Hemanjani Bhavana, Xinxin Peng, Zhongqi Ge, Hu Chen, Jun Li, Zhongyuan Chen, Huiwen Zhang, Leng Han, Di Du, Chad J. Creighton, Gordon B. Mills, Samantha J. Caesar-Johnson, John A. Demchok, Ina Felau, Melpomeni Kasapi, Martin L. Ferguson, Carolyn M. Hutter, Heidi J. Sofia, Roy Tarnuzzer, Zhining Wang, Liming Yang, Jean C. Zenklusen, Jiashan (Julia) Zhang, Sudha Chudamani, Jia Liu, Laxmi Lolla, Rashi Naresh, Todd Pihl, Qiang Sun, Yunhu Wan, Ye Wu, Juok Cho, Timothy DeFreitas, Scott Frazer, Nils Gehlenborg, Gad Getz, David I. Heiman, Jaegil Kim, Michael S. Lawrence, Pei Lin, Sam Meier, Michael S. Noble, Gordon Saksena, Doug Voet, Hailei Zhang, Brady Bernard, Nyasha Chambwe, Varsha Dhankani, Theo Knijnenburg, Roger Kramer, Kalle Leinonen, Yuexin Liu, Michael Miller, Sheila Reynolds, Ilya Shmulevich, Vesteinn Thorsson, Wei Zhang, Rehan Akbani, Bradley M. Broom, Apurva M. Hegde, Zhenlin Ju, Rupa S. Kanchi, Anil Korkut, Jun Li, Han Liang, Shiyun Ling, Wenbin Liu, Yiling Lu, Gordon B. Mills, Kwok-Shing Ng, Arvind Rao, Michael Ryan, Jing Wang, John N. Weinstein, Jiexin Zhang, Adam Abeshouse, Joshua Armenia, Debyani Chakravarty, Walid K. Chatila, Ino de Bruijn, Jianjiong Gao, Benjamin E. Gross, Zachary J. Heins, Ritika Kundra, Konnor La, Marc Ladanyi, Augustin Luna, Moriah G. Nissan, Angelica Ochoa, Sarah M. Phillips, Ed Reznik, Francisco Sanchez-Vega, Chris Sander, Nikolaus Schultz,

Robert Sheridan, S. Onur Sumer, Yichao Sun, Barry S. Taylor, Jioajiao Wang, Hongxin Zhang, Pavana Anur, Myron Peto, Paul Spellman, Christopher Benz, Joshua M. Stuart, Christopher K. Wong, Christina Yau, D. Neil Hayes, Joel S. Parker, Matthew D. Wilkerson, Adrian Ally, Miruna Balasundaram, Reanne Bowlby, Denise Brooks, Rebecca Carlsen, Eric Chuah, Noreen Dhalla, Robert Holt, Steven J. M. Jones, Katayoon Kasaian, Darlene Lee, Yussanne Ma, Marco A. Marra, Michael Mayo, Richard A. Moore, Andrew J. Mungall, Karen Mungall, A. Gordon Robertson, Sara Sadeghi, Jacqueline E. Schein, Payal Sipahimalani, Angela Tam, Nina Thiessen, Kane Tse, Tina Wong, Ashton C. Berger, Rameen Beroukhim, Andrew D. Cherniack, Carrie Cibulskis, Stacey B. Gabriel, Galen F. Gao, Gavin Ha, Matthew Meyerson, Steven E. Schumacher, Juliann Shih, Melanie H. Kucherlapati, Raju S. Kucherlapati, Stephen Baylin, Leslie Cope, Ludmila Danilova, Moiz S. Bootwalla, Phillip H. Lai, Dennis T. Maglinte, David J. Van Den Berg, Daniel J. Weisenberger, J. Todd Auman, Saianand Balu, Tom Bodenheimer, Cheng Fan, Katherine A. Hoadley, Alan P. Hoyle, Stuart R. Jefferys, Corbin D. Jones, Shaowu Meng, Piotr A. Mieczkowski, Lisle E. Mose, Amy H. Perou, Charles M. Perou, Jeffrey Roach, Yan Shi, Janae V. Simons, Tara Skelly, Matthew G. Soloway, Donghui Tan, Umadevi Veluvolu, Huihui Fan, Toshinori Hinoue, Peter W. Laird, Hui Shen, Wanding Zhou, Michelle Bellair, Kyle Chang, Kyle Covington, Chad J. Creighton, Huyen Dinh, HarshaVardhan Doddapaneni, Lawrence A. Donehower, Jennifer Drummond, Richard A. Gibbs, Robert Glenn, Walker Hale, Yi Han, Jianhong Hu, Viktoriya Korchina, Sandra Lee, Lora Lewis, Wei Li, Xiuping Liu, Margaret Morgan, Donna Morton, Donna Muzny, Jireh Santibanez, Margi Sheth, Eve Shinbrot, Linghua Wang, Min Wang, David A. Wheeler, Liu Xi, Fengmei Zhao, Julian Hess, Elizabeth L. Appelbaum, Matthew Bailey, Matthew G. Cordes, Li Ding, Catrina C. Fronick, Lucinda A. Fulton, Robert S. Fulton, Cyriac Kandoth, Elaine R. Mardis, Michael D. McLellan, Christopher A. Miller, Heather K. Schmidt, Richard K. Wilson, Daniel Crain, Erin Curley, Johanna Gardner, Kevin Lau, David Mallery, Scott Mor-

ris, Joseph Paulauskis, Robert Penny, Candace Shelton, Troy Shelton, Mark Sherman, Eric Thompson, Peggy Yena, Jay Bowen, Julie M. Gastier-Foster, Mark Gerken, Kristen M. Leraas, Tara M. Lichtenberg, Nilsa C. Ramirez, Lisa Wise, Erik Zmuda, Niall Corcoran, Tony Costello, Christopher Hovens, Andre L. Carvalho, Ana C. de Carvalho, José H. Fregnani, Adhemar Longatto-Filho, Rui M. Reis, Cristovam Scapulatempo-Neto, Henrique C. S. Silveira, Daniel O. Vidal, Andrew Burnette, Jennifer Eschbacher, Beth Hermes, Ardene Noss, Rosy Singh, Matthew L. Anderson, Patricia D. Castro, Michael Ittmann, David Huntsman, Bernard Kohl, Xuan Le, Richard Thorp, Chris Andry, Elizabeth R. Duffy, Vladimir Lyadov, Oxana Paklina, Galiya Setdikova, Alexey Shabunin, Mikhail Tavobilov, Christopher McPherson, Ronald Warnick, Ross Berkowitz, Daniel Cramer, Colleen Feltmate, Neil Horowitz, Adam Kibel, Michael Muto, Chandrajit P. Raut, Andrei Malykh, Jill S. Barnholtz-Sloan, Wendi Barrett, Karen Devine, Jordonna Fulop, Quinn T. Ostrom, Kristen Shimmel, Yingli Wolinsky, Andrew E. Sloan, Agostino De Rose, Felice Giuliante, Marc Goodman, Beth Y. Karlan, Curt H. Hagedorn, John Eckman, Jodi Harr, Jerome Myers, Kelinda Tucker, Leigh Anne Zach, Brenda Deyarmin, Hai Hu, Leonid Kvecher, Caroline Larson, Richard J. Mural, Stella Somiari, Ales Vicha, Tomas Zelinka, Joseph Bennett, Mary Iacocca, Brenda Rabeno, Patricia Swanson, Mathieu Latour, Louis Lacombe, Bernard Têtu, Alain Bergeron, Mary McGraw, Susan M. Staugaitis, John Chabot, Hanina Hibshoosh, Antonia Sepulveda, Tao Su, Timothy Wang, Olga Potapova, Olga Voronina, Laurence Desjardins, Odette Mariani, Sergio Roman-Roman, Xavier Sastre, Marc-Henri Stern, Feixiong Cheng, Sabina Signoretti, Andrew Berchuck, Darell Bigner, Eric Lipp, Jeffrey Marks, Shannon McCall, Roger McLendon, Angeles Secord, Alexis Sharp, Madhusmita Behera, Daniel J. Brat, Amy Chen, Keith Delman, Seth Force, Fadlo Khuri, Kelly Magliocca, Shishir Maithel, Jeffrey J. Olson, Taofeek Owonikoko, Alan Pickens, Suresh Ramalingam, Dong M. Shin, Gabriel Sica, Erwin G. Van Meir, Hongzheng Zhang, Wil Eijckenboom, Ad Gillis, Esther Korpershoek, Leendert Looijenga, Wolter Oosterhuis, Hans

Stoop, Kim E. van Kessel, Ellen C. Zwarthoff, Chiara Calatozzolo, Lucia Cuppini, Stefania Cuzzubbo, Francesco DiMeco, Gaetano Finocchiaro, Luca Mattei, Alessandro Perin, Bianca Pollo, Chu Chen, John Houck, Pawadee Lohavanichbutr, Arndt Hartmann, Christine Stoehr, Robert Stoehr, Helge Taubert, Sven Wach, Bernd Wullich, Witold Kycler, Dawid Murawa, Maciej Wiznerowicz, Ki Chung, W. Jeffrey Edenfield, Julie Martin, Eric Baudin, Glenn Bubley, Raphael Bueno, Assunta De Rienzo, William G. Richards, Steven Kalkanis, Tom Mikkelsen, Houtan Noushmehr, Lisa Scarpace, Nicolas Girard, Marta Aymerich, Elias Campo, Eva Giné, Armando López Guillermo, Nguyen Van Bang, Phan Thi Hanh, Bui Duc Phu, Yufang Tang, Howard Colman, Kimberley Evason, Peter R. Dottino, John A. Martignetti, Hani Gabra, Hartmut Juhl, Teniola Akeredolu, Serghei Stepa, Dave Hoon, Keunsoo Ahn, Koo Jeong Kang, Felix Beuschlein, Anne Breggia, Michael Birrer, Debra Bell, Mitesh Borad, Alan H. Bryce, Erik Castle, Vishal Chandan, John Cheville, John A. Copland, Michael Farnell, Thomas Flotte, Nasra Giama, Thai Ho, Michael Kendrick, Jean-Pierre Kocher, Karla Kopp, Catherine Moser, David Nagorney, Daniel O'Brien, Brian Patrick O'Neill, Tushar Patel, Gloria Petersen, Florencia Que, Michael Rivera, Lewis Roberts, Robert Smallridge, Thomas Smyrk, Melissa Stanton, R. Houston Thompson, Michael Torbenson, Ju Dong Yang, Lizhi Zhang, Fadi Brimo, Jaffer A. Ajani, Ana Maria Angulo Gonzalez, Carmen Behrens, Jolanta Bondaruk, Russell Broaddus, Bogdan Czerniak, Bita Esmaeli, Junya Fujimoto, Jeffrey Gershenwald, Charles Guo, Alexander J. Lazar, Christopher Logothetis, Funda Meric-Bernstam, Cesar Moran, Lois Ramondetta, David Rice, Anil Sood, Pheroze Tamboli, Timothy Thompson, Patricia Troncoso, Anne Tsao, Ignacio Wistuba, Candace Carter, Lauren Haydu, Peter Hersey, Valerie Jakrot, Hojabr Kakavand, Richard Kefford, Kenneth Lee, Georgina Long, Graham Mann, Michael Quinn, Robyn Saw, Richard Scolyer, Kerwin Shannon, Andrew Spillane, Jonathan Stretch, Maria Synott, John Thompson, James Wilmott, Hikmat Al-Ahmadie, Timothy A. Chan, Ronald Ghossein, Anuradha Gopalan, Douglas A. Levine,

Victor Reuter, Samuel Singer, Bhuvanesh Singh, Nguyen Viet Tien, Thomas Broudy, Cyrus Mirsaidi, Praveen Nair, Paul Drwiega, Judy Miller, Jennifer Smith, Howard Zaren, Joong-Won Park, Nguyen Phi Hung, Electron Kebebew, W. Marston Linehan, Adam R. Metwalli, Karel Pacak, Peter A. Pinto, Mark Schiffman, Laura S. Schmidt, Cathy D. Vocke, Nicolas Wentzensen, Robert Worrell, Hannah Yang, Marc Moncrieff, Chandra Goparaju, Jonathan Melamed, Harvey Pass, Natalia Botnariuc, Irina Caraman, Mircea Cernat, Inga Chemencedji, Adrian Clipca, Serghei Doruc, Ghenadie Gorincioi, Sergiu Mura, Maria Pirtac, Irina Stancul, Diana Tcaciuc, Monique Albert, Iakovina Alexopoulou, Angel Arnaout, John Bartlett, Jay Engel, Sebastien Gilbert, Jeremy Parfitt, Harman Sekhon, George Thomas, Doris M. Rassl, Robert C. Rintoul, Carlo Bifulco, Raina Tamakawa, Walter Urba, Nicholas Hayward, Henri Timmers, Anna Antenucci, Francesco Facciolo, Gianluca Grazi, Mirella Marino, Roberta Merola, Ronald de Krijger, Anne-Paule Gimenez-Roqueplo, Alain Piché, Simone Chevalier, Ginette McKercher, Kivanc Birsoy, Gene Barnett, Cathy Brewer, Carol Farver, Theresa Naska, Nathan A. Pennell, Daniel Raymond, Cathy Schilero, Kathy Smolenski, Felicia Williams, Carl Morrison, Jeffrey A. Borgia, Michael J. Liptay, Mark Pool, Christopher W. Seder, Kerstin Junker, Larsson Omberg, Mikhail Dinkin, George Manikhas, Domenico Alvaro, Maria Consiglia Bragazzi, Vincenzo Cardinale, Guido Carpino, Eugenio Gaudio, David Chesla, Sandra Cottingham, Michael Dubina, Fedor Moiseenko, Renumathy Dhanasekaran, Karl-Friedrich Becker, Klaus-Peter Janssen, Julia Slotta-Huspenina, Mohamed H. Abdel-Rahman, Dina Aziz, Sue Bell, Colleen M. Cebulla, Amy Davis, Rebecca Duell, J. Bradley Elder, Joe Hilty, Bahavna Kumar, James Lang, Norman L. Lehman, Randy Mandt, Phuong Nguyen, Robert Pilarski, Karan Rai, Lynn Schoenfield, Kelly Senecal, Paul Wakely, Paul Hansen, Ronald Lechan, James Powers, Arthur Tischler, William E. Grizzle, Katherine C. Sexton, Alison Kastl, Joel Henderson, Sima Porten, Jens Waldmann, Martin Fassnacht, Sylvia L. Asa, Dirk Schadendorf, Marta Couce, Markus Graefen, Hartwig Huland, Guido Sauter,

Thorsten Schlomm, Ronald Simon, Pierre Tennstedt, Oluwole Olabode, Mark Nelson, Oliver Bathe, Peter R. Carroll, June M. Chan, Philip Disaia, Pat Glenn, Robin K. Kelley, Charles N. Landen, Joanna Phillips, Michael Prados, Jeffry Simko, Karen Smith-McCune, Scott VandenBerg, Kevin Roggin, Ashley Fehrenbach, Ady Kendler, Suzanne Sifri, Ruth Steele, Antonio Jimeno, Francis Carey, Ian Forgie, Massimo Mannelli, Michael Carney, Brenda Hernandez, Benito Campos, Christel Herold-Mende, Christin Jungk, Andreas Unterberg, Andreas von Deimling, Aaron Bossler, Joseph Galbraith, Laura Jacobus, Michael Knudson, Tina Knutson, Deqin Ma, Mohammed Milhem, Rita Sigmund, Andrew K. Godwin, Rashna Madan, Howard G. Rosenthal, Clement Adebamowo, Sally N. Adebamowo, Alex Boussioutas, David Beer, Thomas Giordano, Anne-Marie Mes-Masson, Fred Saad, Therese Bocklage, Lisa Landrum, Robert Mannel, Kathleen Moore, Katherine Moxley, Russel Postier, Joan Walker, Rosemary Zuna, Michael Feldman, Federico Valdivieso, Rajiv Dhir, James Luketich, Edna M. Mora Pinero, Mario Quintero-Aguilo, Carlos Gilberto Carlotti, Jose Sebastião Dos Santos, Rafael Kemp, Ajith Sankarankuty, Daniela Tirapelli, James Catto, Kathy Agnew, Elizabeth Swisher, Jenette Creaney, Bruce Robinson, Carl Simon Shelley, Eryn M. Godwin, Sara Kendall, Cassaundra Shipman, Carol Bradford, Thomas Carey, Andrea Haddad, Jeffey Moyer, Lisa Peterson, Mark Prince, Laura Rozek, Gregory Wolf, Rayleen Bowman, Kwun M. Fong, Ian Yang, Robert Korst, W. Kimryn Rathmell, J. Leigh Fantacone-Campbell, Jeffrey A. Hooke, Albert J. Kovatich, Craig D. Shriver, John DiPersio, Bettina Drake, Ramaswamy Govindan, Sharon Heath, Timothy Ley, Brian Van Tine, Peter Westervelt, Mark A. Rubin, Jung Il Lee, Natália D. Aredes, Armaz Mariamidze, Fernando Camargo, and Han Liang. Comprehensive Molecular Characterization of the Hippo Signaling Pathway in Cancer. *Cell Reports*, 25(5):1304–1317.e5, October 2018.

[169] Larissa Mourao, Guillaume Jacquemin, Mathilde Huyghe, Wojciech J. Nawrocki, Naoual Menssouri, Nicolas Servant, and Silvia Fre. Lineage

tracing of Notch1-expressing cells in intestinal tumours reveals a distinct population of cancer stem cells. *Scientific Reports*, 9(1):888, January 2019. Number: 1 Publisher: Nature Publishing Group.

[170] Clara Morral, Jelena Stanisavljevic, Xavier Hernando-Momblona, Elisabetta Mereu, Adrián Álvarez Varela, Carme Cortina, Diana Stork, Felipe Slebe, Gemma Turon, Gavin Whissell, Marta Sevillano, Anna Merlos-Suárez, Àngela Casanova-Martí, Catia Moutinho, Scott W. Lowe, Lukas E. Dow, Alberto Villanueva, Elena Sancho, Holger Heyn, and Eduard Batlle. Zonation of Ribosomal DNA Transcription Defines a Stem Cell Hierarchy in Colorectal Cancer. *Cell Stem Cell*, 26(6):845–861.e12, June 2020.

[171] Damian Smedley, Syed Haider, Benoit Ballester, Richard Holland, Darin London, Gudmundur Thorisson, and Arek Kasprzyk. BioMart – biological queries made easy. *BMC Genomics*, 10(1):22, January 2009.

[172] Taiyun Wei and Viliam Simko. *R package 'corrplot': Visualization of a Correlation Matrix*. Software Package, 2021.

[173] Indrajeet Patil. statsExpressions: R Package for Tidy Dataframes and Expressions with Statistical Details. *Journal of Open Source Software*, 6(61):3236, 2021. Publisher: The Open Journal.

[174] Tiago P. Peixoto. The graph-tool python library. *figshare*, 2014.

[175] Ken Bellock, Neil Godber, and Philip Kahn. bellockk/alphashape: v1.3.1 Release, April 2021.

[176] Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, C J Carey, İlhan Polat, Yu Feng, Eric W. Moore, Jake VanderPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian

Henriksen, E. A. Quintero, Charles R. Harris, Anne M. Archibald, Antônio H. Ribeiro, Fabian Pedregosa, Paul van Mulbregt, and SciPy 1.0 Contributors. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 17:261–272, 2020.

[177] Suoqin Jin. sqjin/CellChat, January 2021. original-date: 2020-04-27T06:28:33Z.

[178] Robin Browaeys, Wouter Saelens, and Yvan Saeys. NicheNet: modeling intercellular communication by linking ligands to target genes. *Nature Methods*, 17(2):159–162, February 2020. Number: 2 Publisher: Nature Publishing Group.

[179] Aric A. Hagberg, Daniel A. Schult, and Pieter J. Swart. Exploring Network Structure, Dynamics, and Function using NetworkX. In Gaël Varoquaux, Travis Vaught, and Jarrod Millman, editors, *Proceedings of the 7th Python in Science Conference*, pages 11 – 15, Pasadena, CA USA, 2008.

[180] Zhenghang Zhang, Jinlu Jia, Yalin Wan, Yang Zhou, Yuting Kong, Yurong Qian, and Jun Long. TransR *: Representation learning model by flexible translation and relation matrix projection. *Journal of Intelligent & Fuzzy Systems*, 40(5):10251–10259, 2021. Publisher: IOS Press.

[181] Mehdi Ali, Max Berrendorf, Charles Tapley Hoyt, Laurent Vermue, Sahand Sharifzadeh, Volker Tresp, and Jens Lehmann. PyKEEN 1.0: A Python Library for Training and Evaluating Knowledge Graph Embeddings. *Journal of Machine Learning Research*, 22(82):1–6, 2021.

[182] Antoine Bordes, Nicolas Usunier, Alberto Garcia-Duran, Jason Weston, and Oksana Yakhnenko. Translating Embeddings for Modeling Multi-relational Data. In *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc., 2013.

[183] Thomas M. J. Fruchterman and Edward M. Reingold. Graph drawing by force-directed placement. *Software: Prac-*

tice and Experience, 21(11):1129–1164, 1991. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/spe.4380211102.

[184] Greg Turk and Marc Levoy. Zippered Polygon Meshes from Range Images. In *Proceedings of the 21st Annual Conference on Computer Graphics and Interactive Techniques*, SIGGRAPH '94, pages 311–318, New York, NY, USA, 1994. Association for Computing Machinery.

[185] Charles R. Harris, K. Jarrod Millman, Stéfan J. van der Walt, Ralf Gommers, Pauli Virtanen, David Cournapeau, Eric Wieser, Julian Taylor, Sebastian Berg, Nathaniel J. Smith, Robert Kern, Matti Picus, Stephan Hoyer, Marten H. van Kerkwijk, Matthew Brett, Allan Haldane, Jaime Fernández del Río, Mark Wiebe, Pearu Peterson, Pierre Gérard-Marchant, Kevin Sheppard, Tyler Reddy, Warren Weckesser, Hameer Abbasi, Christoph Gohlke, and Travis E. Oliphant. Array programming with NumPy. *Nature*, 585(7825):357–362, September 2020. Publisher: Springer Science and Business Media LLC.

[186] Ilaria M. Michelozzi, Eduardo Gomez-Castaneda, Ruben V. C. Pohle, Ferran Cardoso Rodriguez, Jahangir Sufi, Pau Puigdevall Costa, Meera Subramaniyam, Efstratios Kirtsios, Ayad Eddaoudi, Si Wei Wu, Aleks Guvenel, Jonathan Fisher, Sara Ghorashian, Martin A. Pule, Christopher J. Tape, Sergi Castellano, Persis J. Amrolia, and Alice Giustacchini. Activation priming and cytokine polyfunctionality modulate the enhanced functionality of low-affinity CD19 CAR T cells. *Blood Advances*, 7(9):1725–1738, April 2023.

[187] Plotly Plotly Technologies Inc. Collaborative data science, 2015. Place: Montreal, QC Publisher: Plotly Technologies Inc.

[188] Ron H. J. Mathijssen, Robbert J. van Alphen, Jaap Verweij, Walter J. Loos, Kees Nooter, Gerrit Stoter, and Alex Sparreboom. Clinical Pharmacokinetics and Metabolism of Irinotecan (CPT-11). *Clinical Cancer Research*, 7(8):2182–2194, August 2001.

[189] Manolis Roulis and Richard A. Flavell. Fibroblasts and myofibroblasts of the intestinal lamina propria in physiology and disease. *Differentiation*, 92(3):116–131, September 2016.

[190] Raymond A. Isidro and Caroline B. Appleyard. Colonic macrophage polarization in homeostasis, inflammation, and cancer. *American Journal of Physiology - Gastrointestinal and Liver Physiology*, 311(1):G59–G73, July 2016.

[191] Maxime M. Mahe, Eitaro Aihara, Michael A. Schumacher, Yana Zavros, Marshall H. Montrose, Michael A. Helmrath, Toshiro Sato, and Noah F. Shroyer. Establishment of Gastrointestinal Epithelial Organoids. *Current Protocols in Mouse Biology*, 3(4):217–240, 2013. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/9780470942390.mo130179.

[192] Olga N. Karpus, B. Florien Westendorp, Jacqueline L. M. Vermeulen, Sander Meisner, Jan Koster, Vanesa Muncan, Manon E. Wildenberg, and Gijs R. van den Brink. Colonic CD90+ Crypt Fibroblasts Secrete Semaphorins to Support Epithelial Growth. *Cell Reports*, 26(13):3698–3708.e5, March 2019.

[193] Yuji Naito, Tomohisa Takagi, and Yasuki Higashimura. Heme oxygenase-1 and anti-inflammatory M2 macrophages. *Archives of Biochemistry and Biophysics*, 564:83–88, December 2014.

[194] John T. Hancock. 3 - The principles of cell signalling. In Sanjeev Kumar and Peter J. Bentley, editors, *On Growth, Form and Computers*, pages 64–81. Academic Press, London, 2003.

[195] Wei ZHANG and Hui Tu LIU. MAPK signal pathways in the regulation of cell proliferation in mammalian cells. *Cell Research*, 12(1):9–18, March 2002.

[196] Aditya Pratapa, Amogh P. Jalihal, Jeffrey N. Law, Aditya Bharadwaj, and T. M. Murali. Benchmarking algorithms for gene regulatory network inference from

single-cell transcriptomic data. *Nature Methods*, 17(2):147–154, February 2020.

[197] David S. Fischer, Anna C. Schaar, and Fabian J. Theis. Modeling intercellular communication in tissues using spatial graphs of cells. *Nature Biotechnology*, pages 1–5, October 2022. Publisher: Nature Publishing Group.

[198] Yongjian Yang, Guanxun Li, Yan Zhong, Qian Xu, Yu-Te Lin, Cristhian Roman-Vicharra, Robert S. Chapkin, and James J. Cai. scTenifoldXct: A semi-supervised method for predicting cell-cell interactions and mapping cellular communication graphs. *Cell Systems*, 14(4):302–311.e4, April 2023. Publisher: Elsevier.

[199] Ines Chami, Rex Ying, Christopher Ré, and Jure Leskovec. Hyperbolic Graph Convolutional Neural Networks, October 2019. arXiv:1910.12933 [cs, stat].

[200] Igor Stzepourginski, Giulia Nigro, Jean-Marie Jacob, Sophie Dulauroy, Philippe J. Sansonetti, Gérard Eberl, and Lucie Peduto. CD34+ mesenchymal cells are a major component of the intestinal stem cells niche at homeostasis and after injury. *Proceedings of the National Academy of Sciences*, 114(4):E506–E513, January 2017. Publisher: Proceedings of the National Academy of Sciences.

[201] Jianxi Luo and Christopher L. Magee. Detecting evolving patterns of self-organizing networks by flow hierarchy measurement. *Complexity*, 16(6):53–61, 2011. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/cplx.20368.