Tumorigenic Metaplasia in Colorectal Cancer Delineated Through Applied Systems Theory

By

Bob Chen

Dissertation

Submitted to the Faculty of the

Graduate School of Vanderbilt University

In partial fulfillment of the requirements

For the degree of

DOCTOR OF PHILOSOPHY

In

Chemical and Physical Biology

February 28, 2022

Nashville, TN

Approved:

Vito Quaranta, M.D., Ph.D.

Emily Hodges, Ph.D.

Qi Liu, Ph.D.

Mark N. Ellingham, Ph.D.

Christopher S. Williams, M.D. Ph.D.

Ken. S. Lau, Ph.D.

Dedication

Nanos gigantium humeris insidentes

Acknowledgements

My mentors at Vanderbilt were pivotal in my doctoral training, and my accomplishments would not have been possible without their exceptional guidance. I could not have asked for more in a doctoral advisor, Dr. Ken Lau, and am tremendously grateful for his tireless dedication to my training, student advocacy, and the scientific process. Dr. Emily Hodges was the first to introduce me to the caliber of research done at this institution. She set me up for success at Vanderbilt, and I am grateful that she continued to advise me as a member of my doctoral committee alongside Dr. Vito Quaranta, Dr. Qi Liu, Dr. Chris Williams, and Dr. Mark Ellingham.

I am also grateful for the first-generation of Ken's lab members: Chuck for his sardonic wit and his establishment of this group's computational expertise, Amrita for her academic wisdom and insights about graduate school, Cherie' for her energetic rapport and deep knowledge of cancer, and Joey for his positivity, spontaneous creativity, and extraordinary understanding of how all the lab's moving parts fit together.

Special acknowledgement is deserved by several members of the Department of Biomedical Informatics, who provided a significant amount of mentorship and funding (Big Biomedical Data Science Training Program T32LM012412) for my graduate training while enabling my incursions into medical informatics research: Dr. You Chen, Dr. Brad Malin, and Dr. Cindy Gadd.

Finally, I am grateful for the long list of close friends I have been blessed with, who have supported me throughout my time at Vanderbilt: Krish, Borg, Kirill, Lucas, Alex, Kevin, David, Harrison, Eric, Jacob, Christian, Connor, Faraz, Anthony, Johnny, Mary Kate, Jacky, Laura, Jess, Haley, and my brother Luke.

Table of Contents

List of Figures

List of Tables

Preface

This dissertation is structured as five chapters. Chapters I and II are general overviews of background and recreated from published literature, which are referenced when necessary. These chapters serve as an introduction to complex systems, systems biology, single-cell biology, and cancer biology. Chapters III, IV, and V are recreated from the manuscript titled **Differential pre-malignant programs and microenvironment chart distinct paths to malignancy in human colorectal polyps.** Chapter III details the data-driven discovery of divergent tumor-associated regulatory programs. Chapter IV further describes these observed effects within the tumor microenvironment alongside genetically engineered mouse and organoid models used for validation. Chapter V discusses the wider implications of these discoveries and presents potential directions for next-generation methods in the analysis of complex biological systems.

Chapter I - Systems theory and its application to data-driven biomedical science

Recreated from:

Herring, C. A., **Chen, B.**, McKinley, E. T., & Lau, K. S. (2018). Single-Cell Computational Strategies for Lineage Reconstruction in Tissue Systems. *Cellular and Molecular Gastroenterology and Hepatology*. https://doi.org/10.1016/j.jcmgh.2018.01.023

and

Chen, B., Alrifai, W., Gao, C., Jones, B., Novak, L., Lorenzi, N., France, D., Malin, B., & Chen, Y. (2021). Mining tasks and task characteristics from electronic health record audit logs with unsupervised machine learning. *Journal of the American Medical Informatics Association.* https://doi.org/10.1093/jamia/ocaa338

and

Islam, M., **Chen, B.**, Spraggins, J. M., Kelly, R. T., & Lau, K. S. (2020). Use of Single-Cell -Omic Technologies to Study the Gastrointestinal Tract and Diseases, From Single Cell Identities to Patient Features. Gastroenterology. https://doi.org/10.1053/j.gastro.2020.04.073

Complex Systems

Deconstructing observed natural phenomena has historically been approached through bottom-up paradigms. Such paradigms involve the ostensible breakdown of highly complex, macroscale systems into their elemental, molecular components followed by the perturbation and modeling of these components through hypothesis-driven procedures. The investigation of these phenomena begins at the scale of human perception; thus, these observations describe complex systems, abstracted several hierarchical levels away from their underlying molecular mechanisms [1,2]. While a deep understanding of such elements is possible, there remains a disconnect between a collection of well-characterized components and the constituent system's overall functional or

behavior [3,4]. Though it can be intuited what complex systems consist of, they can be generally identified by their shared features with respect to emergence, robustness, and modularity.

Emergent behavior in complex systems is characterized by the observation of properties not directly found in its elemental components. Early models of complex systems, such as cellular automata, were defined by simple rule sets **(Figure 1)** [5]. Given a large enough space of possibilities, it would quickly become unfeasible or even impossible to predict the final state of the system due to emergent behaviors [6,7]. These behaviors are emergent as the participating agents, or components, follow relatively simple sets of rules within the system, but the overall behavior of the system does not directly correspond to the behavior of any individual agents; nor are these behaviors easily predictable [8]. A less theoretical example is that human cells, while also complex systems themselves, act in aggregate to dictate the behavior of an organ. Another more sociologically scaled example is individual healthcare providers acting in concert within a healthcare organization (HCO) [9]. The key observations here are that organs do not appear to be scaled up single-cells nor are HCOs scaled up healthcare providers; this is succinctly described as being “greater than the sum of its parts”.

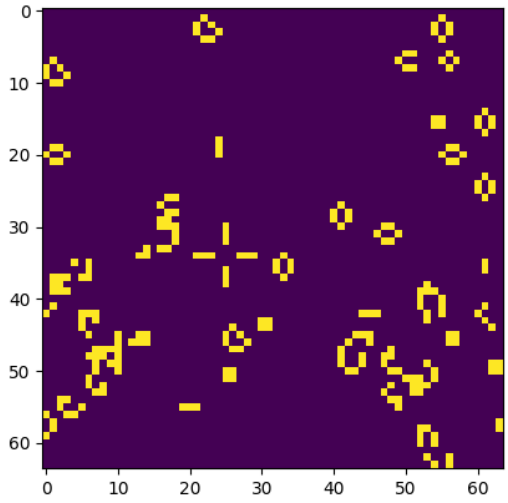


Figure 1. Conway's Game of Life.

An implementation of John Conway's cellular automaton in python, modeling a complex system of 'cells'. This simulation follows simple rules: Given two or three cell neighbors, a living cell remains alive; with three cells giving life to a dead cell. Otherwise, living cells die within one generation and dead cells that do not satisfy these conditions remain unchanged. Though the rules are simple, stable patterns of cells emerge spontaneously until disturbed, such as the four-cell square.

Robust behavior is another feature common to complex systems, meaning that complex systems have a dynamic sensitivity to both internal and external perturbations. In practice, many emergent behaviors can be formalized as the nonlinear amplification or dampening of inputs and outputs, lending to this property of robustness. Feedback loops, for example, consist of simple, self-referential components, and may propagate or dampen signals throughout the system. Reasonably, this tendency to dynamically regulate the flow of information may be necessary for several complex systems to exist at all, given that they are often open systems communicating other systems operating at different scales and modalities. Following the analogy of cells, injury response in organs necessitate processes that regulate the fine-tuned compartmentalization and repair of cellular damage without initiating a cascade of failures throughout the entire organ. Feedback loops are still only one motif that may exist in a network of interacting agents comprising a complex system, which may be symbolically represented as graphs [4,10]. In such network representations, the components of a complex system are graph vertices, and their generic interactions are the edges connecting these vertices.

Modular properties of complex systems accompany these emergent and robust behaviors, where functional network motifs, or systems, may become robust themselves, and thus partially

independent from the operation of the complex system as a whole. Much like how robust systems may emerge from self-organization, its subsystems, too, may undergo a scaled-down version of self-organization. This fractal behavior is also known as scale-free organization, where modules of network components tend to maintain a large number of connections to a small number of important components **(Figure 2)** [11]. The resulting distribution of graph vertices and their degrees of connection, or number of edges, is then observed to follow an asymptotic power law. Assuming random perturbations may occur in an open system, the risk of systemic dysfunction is diluted through all possible vertices and edges, while important components of a complex system remain insulated. This fault-tolerant behavior also permits a relatively safe evolutionary progression of subsystems, as mutations destructive to the network are compartmentalized.

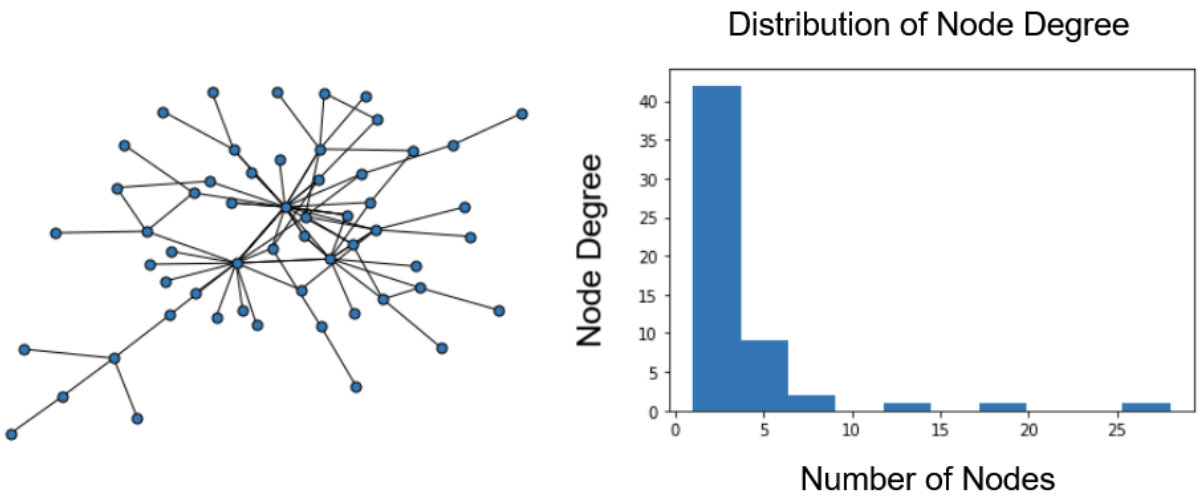


Figure 2. Scale free network.

Graph and respective histogram generated using the Barabási–Albert model in python and the NetworkX package. The distribution of node degree follows a power law.

General Systems Theory

Complex systems are observed across several fields of study with, albeit with different vocabularies and at different scales. General principles outlining such systems were first described in the 1940s and 50s among biologists, physicists, and psychologists among others. Some well-known precursors to General Systems Theory (GST), as introduced by biologist Ludwig von Bertalanffy, include subjects like Gestalt psychology, statistical physics, and ergodic theory for the study of dynamical systems [2]. Bertalanffy and his contemporaries, held that the top-down view of a system and its behavior should be seen as valuable as its elemental components, alongside the generalizability of these types of relationships. Key concepts of GST include the recognition of isomorphic laws across fields of study and its characterization of open systems.

This structural isomorphism of laws across fields refers to the shared concepts that have been discovered in different complex systems, with a simple example being exponential laws and principles of diminishing returns **(Figure 3)**. In nuclear physics and microbiology, radioactive isotopes undergo decay and microbes undergo division; both express diminishing returns as open systems experiencing a net loss of entropy. GST posits that the identification of “generalized kinetics and dynamics” or logically homologous interactions in separate models should lead to a better understanding of complex systems behaviors, especially in the context of top-down approaches. Ultimately, GST acts as a foundational paradigm for interdisciplinary study of complex systems.

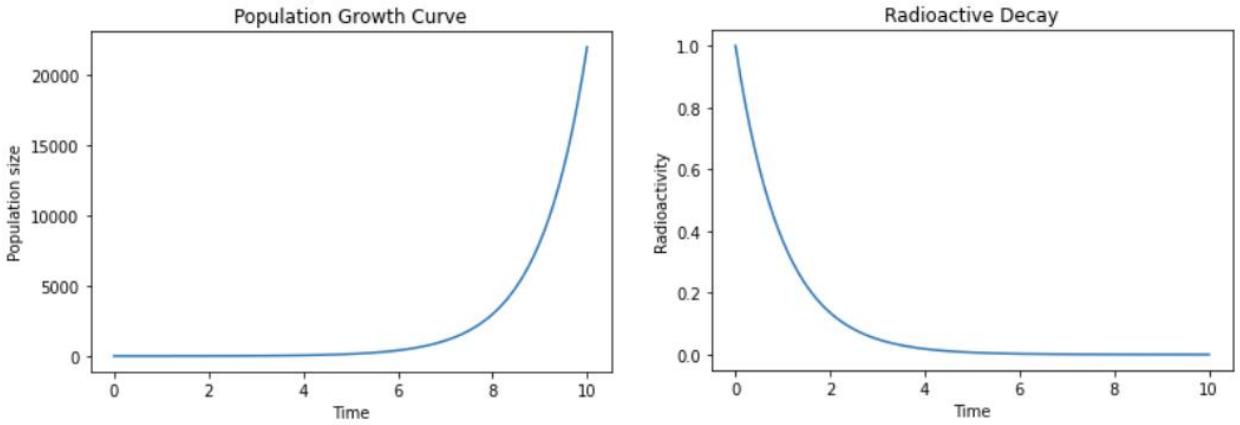


Figure 3. Isomorphic laws of exponential behavior.

A simulated example of exponential behavior between population dynamics and radioactive half-lives using python. An examination of properties such as entropy in either system, for example, may yield insights on diminishing returns or explosive growth.

Cells as Complex Systems

On several hierarchical levels cellular networks are open and complex systems, meaning that the flow of information crosses different scopes and scales of interaction. Primarily, these scales reflect the central dogma of molecular biology, where information freely flows between genomic, transcriptomic, and proteomic layers of molecular organization. Still, these types of interactions are largely intracellular, acting within a single-cell. Intercellular interactions between heterogeneous cell types, such as those comprising organs, are several hierarchical levels abstracted from its underlying molecular mechanics **(Figure 4)**. Modeling and understanding complex systems, such as these, first involves the high-dimensional quantification of its respective features. Given a space of all possible states within a complex system, its cumulative features should be able to approximate a single state. Accordingly, dynamic models would involve the characterization of these state transitions over time.

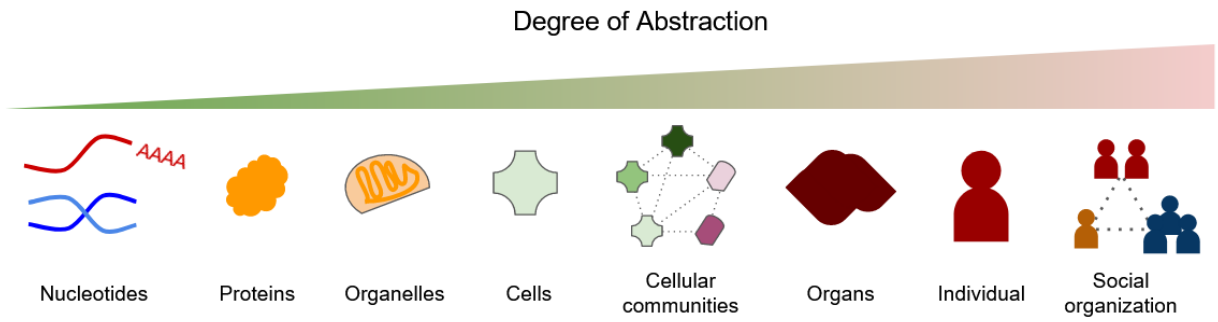


Figure 4. Hierarchical levels of abstraction.

The hierarchical 'grouping' or abstraction of molecular functions compounds into more interdependent complex systems. Moving between these hierarchical levels, it is infeasible to accurately predict the effects of non-adjacent abstractions without experimentation or simulation. For example, it is difficult to infer the effect social structures have on specific organelles and vice versa.

Though modular components of a complex system can be removed and perturbed *in vitro* (i.e. a gene knockout in a mammalian cell line), its altered behavior does not necessarily reflect what may happen in different hierarchies of complex systems. Organ-level behaviors, expectedly, are unfeasible to accurately predict without a complete recapitulation of the system itself, hence the value of *in vivo* experimentation. This disconnect between relatively simple agents and the emergent properties of their overall complex system strongly parallels the computational irreducibility of cellular automata, a concept proposed by Stephen Wolfram on a backdrop of John Conway's 'Game of Life' [7]. Simplistically, computational irreducibility is a property of complex, computable systems where there is no feasible way to predict its final state without performing each procedural step in its given rule set. In the same way that computational fields have benefited from top-down, systems approaches so do complex biological systems.

Omics Paradigm Shifts in Systems Biology

Technological advances have enabled the omics-level molecular profiling of living systems, and more recently at single-cell resolutions. More broadly, this means that high-resolution, top-down characterizations of systems have become significantly more feasible. Systems biology approaches, or the application of systems theory for the understanding of biological systems, have proliferated in conjunction with these new technologies [8,12]. Former technical challenges involving the limited micro-scale tooling and computational resources available have generally been overcome. In line with this, isomorphic examples of Moore's law have been observed both in commercially available computers (measured by the doubling number of transistors) as well as genome sequencing (measured by the exponentially decreasing cost per raw megabase of DNA sequencing) since the beginning of the Human Genome Project (HGP).

Concerted institutional efforts following the HGP, such as the Encyclopedia of DNA Elements (ENCODE) and The Cancer Genome Atlas (TCGA), have resulted in explosions of data on many fronts. The widespread adoption of these technologies represented a preliminary shift away from candidate-based (bottom-up) approaches and towards data-driven (top-down) approaches for hypothesis generation. Notably, methods using massively-parallel sequencing by synthesis were at the core of these projects; as a consequence, nucleotide-based omics protocols allowed for the joint development and wide-spread application of next-generation sequencing devices. The resultant ecosystems of large-scale omics projects then expanded from genomics to transcriptomics, followed by epigenomics and proteomics. With the increasing accessibility of omics methods, the paradigm shifted, again, towards more context-specific processes in biological systems. Functionally, this meant that top-down omics could approach the biological scoping of bottom-up approaches; thus, the abstraction gap between hypothesis-driven and data-driven paradigms narrowed. For example, TCGA aimed to create an integrative map of several human cancer types and their respective tissues by measuring their molecular characteristics, ranging from their gene expression to their DNA methylation profiles. These efforts originated

from the general understanding of the human genome, laid out by the HGP, but refocused on tissue-level (as opposed to patient-level) abstractions of underlying biological phenomena.

These data have been valuable in both basic and translational aspects, but without deeper processing, tissue-level abstractions also have a limited scope with the general assumption cellular homogeneity. In the same way that humans have biological functions compartmentalized to different tissues, these tissues are also compartmentalized into heterogeneous cell types and structures. Tissue-level bulk genomics, therefore, helps to define human biology over populational averages of individual cell types, unable to traverse the layer of cellular complexity **(Figure 5)**. This gap between tissue-level and cell-level complex system abstractions is extremely important since therapeutics ultimately act on the underlying biological mechanisms at the molecular level [13]. The progression of several key technologies, many of which heralded the bulk genomics paradigm, have since brought about the diffusion of single-cell resolution technologies. Outside of the decreasing costs of nucleotide sequencing and strand synthesis (for sequencing library construction), newly developed applications of microfluidic technologies allowed the characterization of single-cells from human tissue.

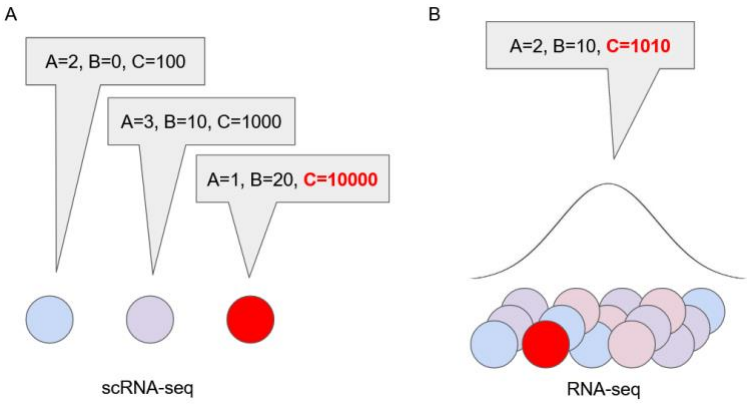


Figure 5. scRNA-seq compared to bulk RNA-seq.

(A) Genes A, B, and C can individually be probed per single cell with scRNA-seq. The cell on the right is expressing 10,000 transcripts of gene C. **(B)** A population average is, instead, measured

for genes A, B, and C; the high expression of gene C in certain cells is diluted through the measurement of the entire population of cells.

Previously, methods such as flow cytometry were used to measure properties of individual cells, such as size and DNA content, through light absorption. Candidate-based approaches applied to these pipelines, by means of fluorescent protein conjugates, allowed for the detection of specific cellular proteins. Combined with microfluidic sorting mechanisms, heterogeneous mixtures of cells, like those found in human tissue, could then be disaggregated by their molecular properties. More recent variations on these candidate-based approaches, like DISSECT-CyTOF, use mass spectrometry to quantify cellular features labeled with heavy metals, which overcome limitations of fluorescent multiplexing such as spectral overlap [14,15]. Advances such as these allow for an increase in the breadth of detectable features in single-cell state characterization. Incremental improvements in candidate-based approaches, while valuable for focusing on specific components of biological pathways, still remain several orders of magnitude away from omics-level feature spaces. Thus, the eventual application of high-throughput sequencing technologies gradually began to replace multiplexed, candidate-based approaches in general top-down characterizations of heterogeneous cell populations.

Single-Cell Transcriptomics

Precursors for commonly used single-cell transcriptomic technologies originated in candidate-based approaches such as microarrays. Typically, large volumes of DNA-based probes are generated per candidate gene target and quantified based on the intensity of its respective hybridization signal. The dynamic range and diversity of genes quantified is limited to those defined at the time of experimentation, but these methods enabled a first step into single-cell resolution transcriptomics. Early adopters, like Yamamura et al. and Kamme et al., combined these microarray technologies with microwells and laser capture [16,17]. Low-throughput adaptations

of RNA-seq at single-cell resolutions, first described by Tang et al., were able to overcome the limitations of these methods, but introduced new challenges in throughput and scalability, having been developed in isolated oocytes. Such challenges were due to the inability of existing technologies to rapidly isolate and deconvolute biologically meaningful volumes of single-cells. Though early developmental models, on the order of eight to twelve cells, significantly benefited, physiologically relevant numbers of cells, on the order of hundreds to thousands, could not be feasibly analyzed.

The maturation of two key technologies, stemming from established high-throughput sequencing methods and microfluidic chip fabrication, facilitated the transition of single-cell transcriptomics into physiologically relevant cell numbers and out of cost-prohibitive protocols [18]. Single-cell droplet encapsulation, and the paradigm shift that followed, traded the burden of scaling cell number from microwells and micro-manipulation onto the reaction reagents themselves, high-throughput sequencing, and computational deconvolution. In this framework, single cells and their barcoded sequencing libraries become completely isolated within a droplet, formed at the interface of oil and an aqueous cell suspension. Generating these droplets involves a tightly controlled flow of a few microfluidic channels containing a dilution of barcoded primer beads and dissociated cells in addition to the reaction mixture and oil interface. Three of the most commonly used platforms, inDrop, Drop-seq, and 10x Chromium are used to generate these emulsifications of droplet-encapsulated cells, each with unique assumptions and drawbacks [19–21].

In terms of modularity and cost-sensitivity, the inDrop platform presents an optimal balance for developing incremental improvements of and rapid application to large-scale human studies. Fundamentally, inDrop features several open-source aspects ranging from microfluidic chip design to its compatibility with a wide range of mechanical components and reagents. Its barcoded beads, containing synthesized nucleotide primer fragments, vary minimally from its

counterparts, Drop-seq and 10x Chromium, consisting of a PCR primer, combinatorically generated cell barcode, unique molecular identifier (UMI), and 3' poly-T tail. Procedurally, each individual droplet reaction produces a reverse-transcribed (RT) transcriptome captured through mRNA poly-A tail annealing and is released from the primer bead through UV photocleavage, an inDrop-specific process. After RT product conversion into double-stranded DNA through second-strand synthesis, a unique feature of the inDrop primer structure, the T7 promoter, is used for in-vitro transcription. This linear amplification process is vital for generating representative, and not exponentially biased, fragments of RNA and its subsequent cDNA product. Finally, adapters suitable for interacting with Illumina sequencing-by-synthesis platforms are appended through PCR, which include P7 and P5 priming sites. The combination of these procedures allows for the high-throughput sequencing of transcribed mRNA with an associated cell barcode and UMI, making each transcriptome tractable to a single cell and each detected transcript tractable to a single mRNA molecule through computational deconvolution, respectively [22].

Single-Cell Genomics and Epigenomics

Whole-transcriptome capture and sequencing only represents one layer of transient, complex molecular information **(Figure 7)**. Looking upstream of the transcriptome, genomic information represents a relatively stable source of all possible transcripts, given other possible post-transcriptional modifications. The intermediary between these two layers of molecular information would then be the epigenome, as a layer of dynamic regulations leading to the flow of information between the genome and transcriptome. Contrasting transcriptomic methods, single-cell genomics and epigenomics have a different set of technical limitations, namely the limited amount of DNA contained within a single nucleus and the physical conformations of its nuclear packaging. Like single-cell transcriptomics, several of these methods are based on the isolation of single cells, the targeted enrichment of nucleotide-based information, the generalized application sequencing, and the use of high-performance computing [23]. Functionally, these methods are often

used for investigating chromatin accessibility, protein–DNA interactions, chromosome conformation, and DNA methylation at single-cell resolutions.

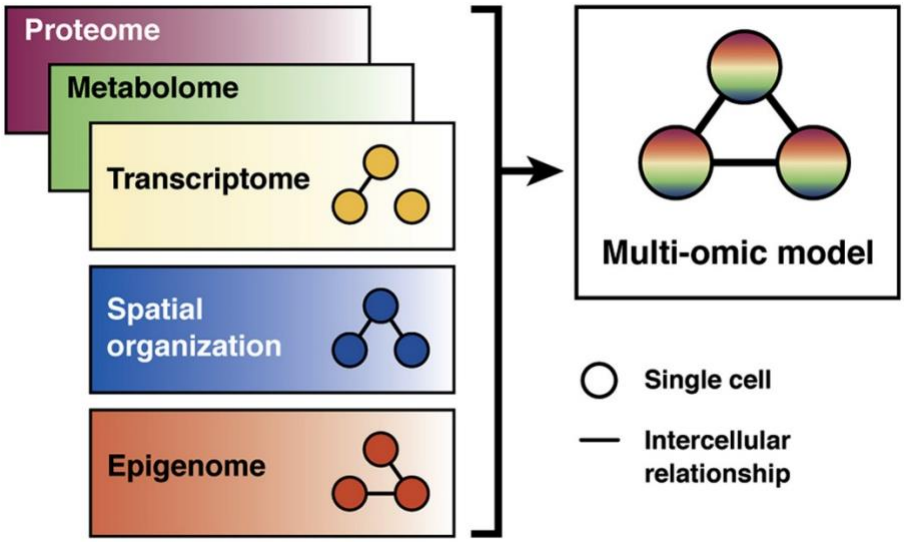


Figure 6. Multi-omic models in single-cell biology.

-Omic level information (proteome, metabolome, transcriptome, spatial organization, epigenome, etc.) may detect cellular relationships individually which are not detected by other modalities. Integrating information across these layers of molecular information yields a complete picture of how these three single-cells are related.

Single-cell assays for transposase-accessible chromatin using sequencing (scATAC-seq) target accessible genomic regions by exploiting the kinetic favorability of Tn5-mediated transposition reactions with DNA not incorporated into nucleosomes. These captured genome sequences can be cis-acting DNA elements poised for transcription or regulation by transcription factors. Borrowing the microfluidic platforms of single-cell transcriptomics, several scATAC-seq methods have been established (see studies by Cusanovich et al, Buenrostro et al, and Lareau et al) [24–26]. These methods isolate individual cells using plated micro-wells, integrated fluidic circuits, and

encapsulation into nanoliter droplets, respectively. Each nucleus therefore produces a single barcoded library of genomic fragments enriched for regions of accessible genomic loci.

Single-cell chromatin conformation capture methods can be used to identify cis-acting DNA elements and determine their physical proximities to potential regulators. Topologically associated domains and long-range chromatin interactions mediated by loop structures can be probed by 3C methods; more recently, these approaches have been advanced to single-cell resolution. Hi-C, and its single-cell variants like sc-Hi-C and sci-Hi-C, developed by Nagano et al and Ramani et al, combine chromatin crosslinking, restriction digestion, and proximity-based ligation to create libraries that capture spatially proximal DNA fragments [27,28]. sci-Hi-C, in particular, isolates nuclei in microwells and incorporates combinatorial indexing. These methods result in a single library per cell, containing fragments that represent pairs of proximally adjacent genomic loci. Although it is not exactly a single-cell chromatin conformation capture method, a modified form of ATAC-seq, called ATAC-See, developed by Chen et al, permits covalent tagging of accessible chromatin with visualizable fluorophores [29]. This allows for visualization by microscopy and subsequent high-throughput sequencing.

Single-cell chromatin immunoprecipitation methods target protein–DNA interactions within single, isolated cells. These methods retain the same strategy as their bulk approaches, relying on specific antibody–protein interactions. Droplet-based single-cell chromatin immunoprecipitation sequencing, a method developed by Rotem et al, takes protein-associated genomic fragments generated from droplet-isolated single cells and tags them with unique DNA barcodes [30]. These nanoliter droplets, which contain the contents of a single cell, are broken and aggregated for immunoprecipitation and library generation. This information can also be obtained using cleavage under targets and tagmentation, described by Kaya-Okur et al [31]. This method uses protein-A–tethered Tn5 transposons to localize these elements to protein-bound antibodies. Target-

localized transposons fragment the genome in a way that enriches for target protein-associated loci. These reactions are amenable to Nanowell-based systems because antibody binding and transposon introduction can be performed at a bulk level prior to isolation. Both of these methods produce sequencing libraries that contain cis-acting regulatory elements associated with the antibody-targeted protein.

Single-cell methylation and hydroxymethylation (sc-5mc and sc-5hmc) assays measure covalent modifications on genomic cytosine residues. Often, these modifications are enriched for CpG islands—high concentrations can result in silencing or reversible down-regulation of gene expression [32]. These methods are classified by their sodium bisulfite dependence, where bisulfite-dependent methods convert cytosine residues into sequencing-detectable uracil. Single-cell genome-wide and reduced-representation sequencing methods, which depend on bisulfite conversion, have been developed to capture varying breadths of the methylome [33,34]. In contrast, single-cell CpG island methylation sequencing combines methylation-sensitive restriction digestions with multiple displacement amplification to generate a sequencing library enriched for loci associated with methylated CpG islands, while avoiding destructive bisulfite conversions [35]. Other new methods include scAba-seq, which targets 5hmc and retains strand-specific information through bisulfite-independent, but glucosylation-dependent enzymatic reactions [36].

Like single-cell transcriptomes, epigenetic data from single cells can be used in human research, possibly to determine patient prognoses and/or to select therapy. Bormann et al. examined the CpG island methylator phenotype along with cell-of-origin signatures in colorectal tumor tissues and identified epigenetically defined subtypes of tumors that correlated with patient survival [37]. Other tumor types have epigenetic heterogeneity along with functional heterogeneity. Litzenburger et al. used scATAC-seq to demonstrate differences in chromatin accessibility associated with sensitivity of cancer cell lines to drugs [38].

Benefiting HCOs Through GST

Machine learning models and applications generalizable to complex systems have been demonstrated through interdisciplinary collaborations between the VUMC (Vanderbilt University Medical Center) Department of Biomedical Informatics and the Chemical and Physical Biology program. In these collaborations, we examined the emergent behaviors of interacting healthcare workers through a digital medium of the Epic electronic health record (EHR) system and in the context of the VUMC Neonatal Intensive Care Unit. Additionally, these methods borrow from methods in the computational analysis of single-cell biology primarily through the encoding of behavioral sessions as analogous to single-cell transcriptomes. Understanding these interactions provides a means of addressing emergent phenomena such as clinician burnout and the quality of patient care [9,39].

Healthcare organizations exist as complex systems of individual providers interacting with each other using a combination of analog and digital means. Clinician activities within EHR systems can influence their workload and workflow, which can induce stress and burnout if improperly managed [40–45]. Clinicians use EHRs for various functions, including chart review, documentation, messaging, orders, patient discovery, medication reconciliation, etc. [46] Healthcare organizations (HCOs) and EHR vendors have previously investigated such usages to understand clinician EHR activities and efficiency [47,48]. These investigations measure the time spent on each EHR function to build provider efficiency profiles [49].

In recent years, EHR audit logs have become valuable resources for the investigation of clinician efficiency in EHRs [50,51]. When a clinician accesses or moves between modules in the EHR interface, such as moving from Progress Notes to Order Entry screens, a timestamped record of that action is documented, along with clinician and patient identifiers [52–55]. Arndt et al. leveraged

audit logs to identify 15 tasks, including clerical (eg, assigning Current Procedural Terminology and International Classification of Diseases–Tenth Revision codes), medical care (eg, reviewing an encounter note), and inbox tasks (eg, developing a letter to a patient) completed by family medicine physicians. Similarly, Sinsky et al. created a set of metrics to quantify the time spent by a physician in an EHR by using audit logs.

Our approach applies unsupervised learning methods to audit log event sequences to identify and characterize putative EHR tasks performed by clinicians. This framework encodes each human-computer interaction (HCI) as feature-rich sessions such that individual functional interactions can be described in a high-dimensional feature space. Relative to the complex task of inpatient care, these behaviors are simple and performed by interacting agents, and such interactions operate at a sociological scale as opposed to the molecular scale observed in single-cells. Structural isomorphisms, however, can be observed between these systems such that the development of complex systems analysis methods may benefit the study of both **(Table 1)**.

Level of Abstraction	**Systems Biology**	**EHR Systems**
Low	Expressed transcripts	Logged HCI events
	Are used to define ...	
	Single cells	User sessions
	Are aggregated and classified as ...	
	Cell types	Provider tasks
	Are mined to understand ...	

High	Biological pathways	Provider behavior trees

Table 1. Isomorphisms between the contextual abstractions of systems biology and EHR systems.

Two contexts of complex systems, in systems biology and EHRs, share several logically analogous

Formally, this framework consists of three hierarchical levels of abstraction, from low to high: the system-level, session-level, and task-levels **(Figure 7)**. The goal of our work is to provide an informatics framework for mining audit log data to discover EHR event and session patterns, which we call tasks. We developed hierarchical metrics to describe these tasks and leveraged them to investigate task complexity and efficiency. System-level representations of the data consist of the most granular observations of clinician-EHR interactions, being the audit log itself. These dense, highly voluminous data are largely uninterpretable, consisting of EHR event categories, provider IDs, patient IDs, and timestamps. We reasoned that the examination of emergent behaviors could only be interpretable at higher levels of abstraction, or the task level.

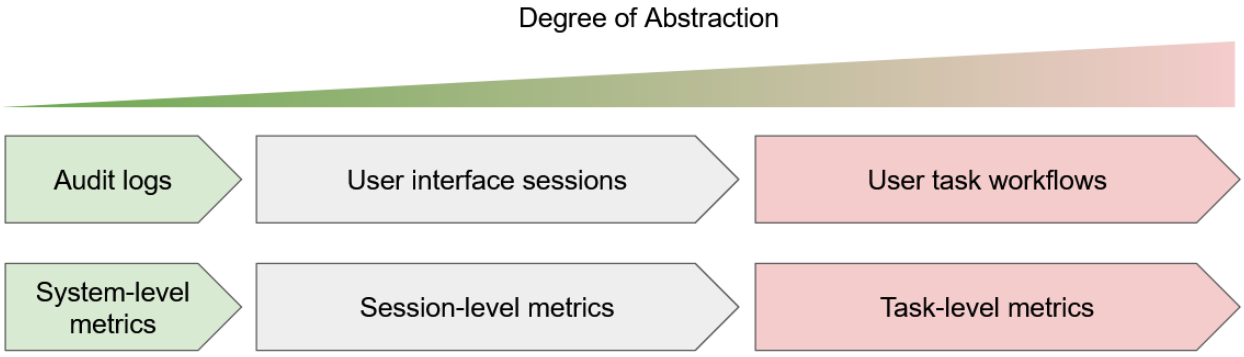


Figure 7. Hierarchical contexts and metrics in EHR analysis.

Each level of abstraction is built upon the previous level, and we devised respective metrics per level. Audit logs represent the raw data in this framework and thus the metrics associated with it

are the baseline at the system-level. Respectively, session and task-level metrics are averaged across their constituent lower-level elements.

To segment these streams of raw data into functionally interpretable abstractions, we used a data-driven method to determine optimal time cutoffs for aggregating HCIs (further elaborated in Chapter II in the context of single-cell computational biology). This was done under the assumption that successive interactions would occur in ergonomic 'bursts' followed by longer pauses **(Figure 8)**. The length of these bursts and pauses were determined through finding the operating point of a cumulative sum of intervals between measured HCIs, which would act as the point of diminishing returns and maximize the density of functional HCI information per session. At this stage, each session is defined by a set of elemental HCIs within a data-driven time window. Such aggregation abstracts the elemental information into a form analogous to a single-cell transcriptome, much like the process of cell encapsulation from heterogeneous tissues.

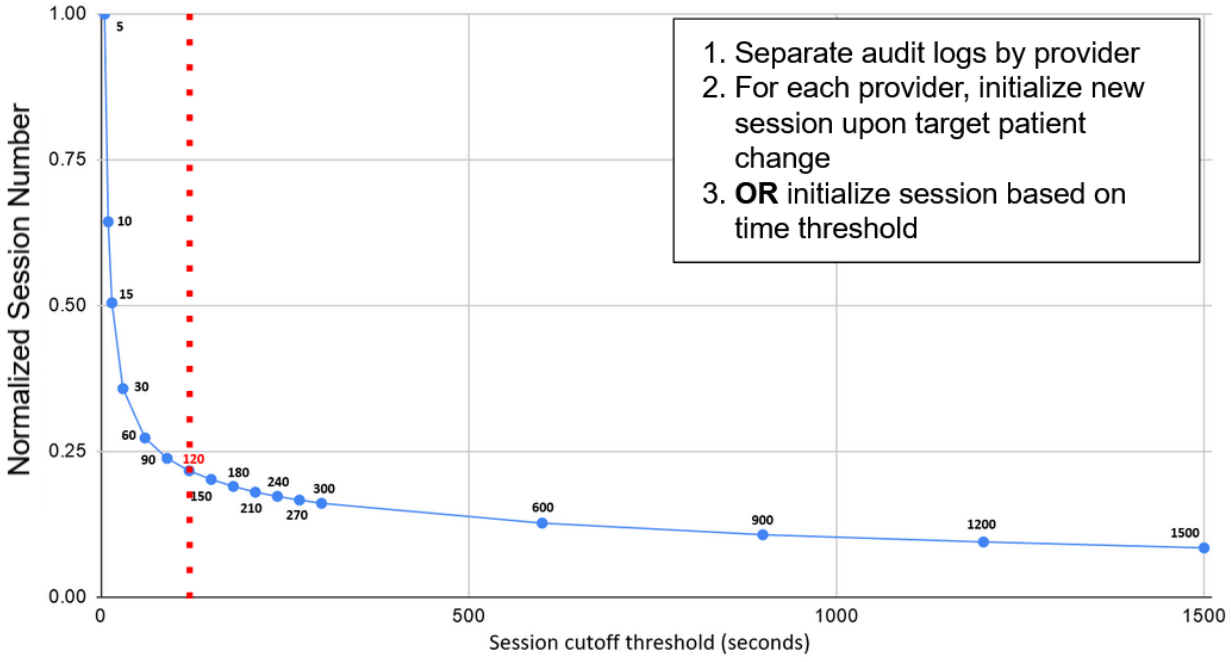


Figure 8. Data-driven audit log sessionization.

Higher-order abstractions are made possible through the transformation and aggregation of lower-level elements. In this case, a continuous list of events is abstracted as a session by finding the operating point of intervals between provider interactions with the EHR system and aggregating those sequences of events.

Because of this session-level abstraction, methods applicable to the analysis of agents defined by a high-dimensional feature space can now be used through the mapping of logical homologies between complex systems. This includes linear dimensionality reduction methods such as principal component analysis (PCA), non-linear projection methods such as Uniform Manifold Approximation and Projection (UMAP) or t-Distributed Stochastic Neighbor Embedding (tSNE), and unsupervised clustering. Primarily, these methods, often used to abstract cellular function in single-cell biology, could instead be used to abstract groups of HCI function in the form of tasks. As a pilot study, this work focuses on task complexity, task efficiency, and task prevalence among clinicians. We stratify tasks by complexity and investigate differences in task efficiency and clinician prevalence between each complexity profile **(Figure 9)**. Our methods can potentially guide HCOs to optimize EHR activities or clinical workflows, by highlighting specific inefficient tasks. To test our methods, we applied our approach to identify and characterize EHR tasks for nurses involved in the care of surgical cases in the neonatal intensive care unit (NICU). Metrics generated relative to each level of abstraction are then used to summarize behavioral information with greater resolution and context-awareness.

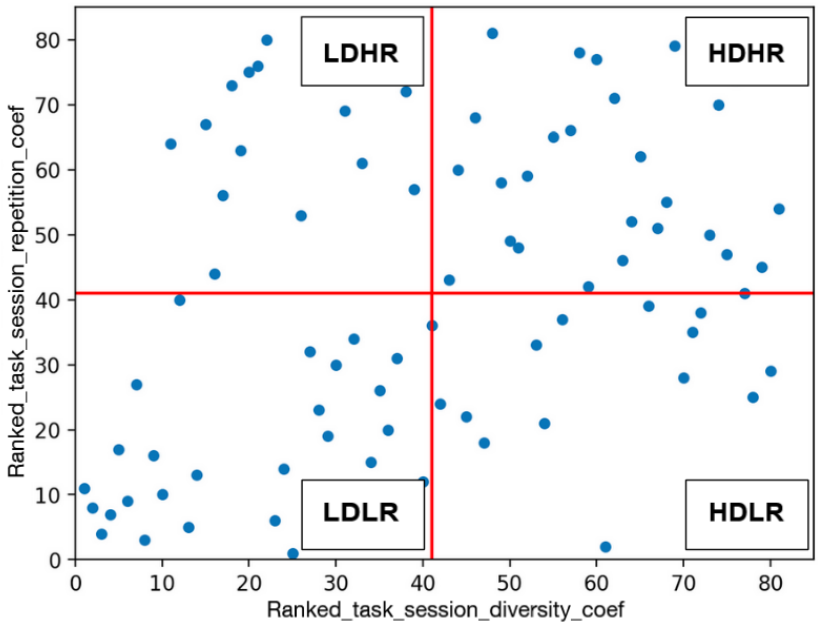


Figure 9. Task complexity profiles.

Each of the 81 detected provider tasks can be split into quadrants of low (L) or high (H) diversity (D) and repetition (R). A ratio was taken between the ranked repetition (how repetitive a task was) and diversity (how diverse the events within a task were) coefficients, and then and split across both axes by the median value.

Studies to date have created metrics to measure the amount of time spent by clinicians in EHRs and the tasks that clinicians perform while interacting with EHR systems, primarily in outpatient and ambulatory settings. Yet few studies have focused on inpatient settings or have developed metrics to model task complexity in EHR utilization. Our study created an unsupervised learning framework to identify clinician tasks performed in EHRs and developed hierarchical metrics to describe EHR task complexity, which is lacking in the existing literature. We tested the effectiveness of our metrics and approach in learning EHR tasks and task complexities for nurses in the NICU. Our hierarchical metrics capture contextual information of a task beyond its explicit, session-level content. Ultimately, by applying the approaches described in this study,

investigators can better understand clinician activities in EHRs with reduced manual effort with the utilization of machine learning.

Chapter II - Single-cell approaches for the investigation of mammalian gastrointestinal biology

Recreated from:

Chen, B., Ramirez, M., Heiser, C., Liu, Q., Lau, & K. S. (2021). Processing single-cell RNA-seq data for dimension reduction-based analyses using open-source tools. *STAR Protocols.* https://doi.org/10.1016/j.xpro.2021.100450

and

Chen, B., Herring, C. A., & Lau, K. S. (2019). pyNVR: investigating factors affecting feature selection from scRNA-seq data for lineage reconstruction. *Bioinformatics*. https://doi.org/10.1093/bioinformatics/bty950

and

Herring, C. A., **Chen, B.**, McKinley, E. T., & Lau, K. S. (2018). Single-Cell Computational Strategies for Lineage Reconstruction in Tissue Systems. *Cellular and Molecular Gastroenterology and Hepatology*. https://doi.org/10.1016/j.jcmgh.2018.01.023

Technical Challenges of Single-Cell Biology and Computational Solutions

Single-cell RNA-sequencing (scRNA-seq) extracts transcriptomic information while preserving complex, multicellular interactions. This is unlike bulk transcriptomic methods, where tissues are homogenized and such cellular heterogeneity is lost in the process, with no feasible way to deconvolve individual cells *de novo*. Single-cell techniques capture extremely complex cell states in the form of high-dimensional data, most often in transcriptomic spaces. scRNA-seq is known to produce noisy data on a per-feature basis, especially for lowly expressed genes, owing to the capture and amplification of small amounts of nucleic acids. These physical limitations are exacerbated by biological phenomenon such as bursting transcription, which entails the stochastic expression and detection of specific transcripts, owing to the fact that scRNA-seq only

captures a temporal snapshot of complex cell states [56]. This effect is compounded with the sparse quantification of expressed features in multidimensional space, which is a phenomenon known as the curse of dimensionality, which greatly affects downstream trajectory analysis when using the full ensemble of features [57].

A way to mitigate this effect is to select and analyze only a subset of the most important features that maximally captures the phenomenon of interest, while ignoring uninformative or noisy features. The feature selection step is implicitly performed in candidate-based approaches, such as Cytometry Time-of-Flight and multiplex microscopy, because the user is picking the most important markers to measure. How to pick informative features while eliminating uninformative ones from genome-scale scRNA-seq experiments is still an active area of research.

One intuitive method for feature selection is a supervised approach that only includes genes of interest. For instance, candidate genes can be selected from a differentially expressed gene set from a bulk RNA-seq experiment that uses a time course or genetic perturbation experimental design. Pipelines such as Single-cell Topological Data Analysis and Single Cell Lineage Inference Using Cell Expression Similarity and Entropy incorporate annotated gene sets from gene ontology resources such as Protein ANalysis THrough Evolutionary Relationships or the Database for Annotation, Visualization and Integrated Discovery to select features in a semi-supervised fashion [58,59]. For studies with minimal or unreliable prior knowledge, completely unsupervised methods that leverage general gene expression patterns may be used.

Different unsupervised feature selection methods vary in their assumptions as well as complexity. For example, a commonly used method in analyzing scRNA-seq data involves identifying transcriptomic features with highly variable expression across the entire data set of single cells. Here, the assumption is that variance in gene expression between cells corresponds to

meaningful gene regulation. This method calculates the variance of each gene across all data points (cells), and filters the features to capture only those with the highest variances [60]. In a way, this method is analogous to principal component analysis (PCA) in selecting the dimensions with the highest variances [61]. Technical variation can potentially exceed meaningful biological variation, and filtering methods can be confounded by the simultaneous occurrence of these 2 sources of variation. However, because of their computational tractability, variance ranking methods can provide a quick evaluation of data quality by enumerating the number of biologically relevant genes returned, which can be collected to potentially reveal both known and unknown cellular relationships.

More sophisticated methods based on different patterns of gene expression have been developed to identify biologically relevant features. Qui et al. developed dpFeature, a method that selects differentially expressed genes between cell populations described by unsupervised clustering for downstream trajectory analysis [62]. Clusters of cells automatically identified are representative of distinct cell states, and differentially expressed genes represent likely regulators of these states. However, data sets that depict transitions are generally continuously distributed and do not form distinct clusters. Clustering in these cases is based on arbitrary cut-off values, and, thus, how dpFeature performs on these types of data sets remains to be tested.

To handle continuous data distributions, Welch et al. developed a metric called neighborhood variance [63]. Implementing a K-nearest neighbors graph approach with each cell represented as a node, this method defines neighborhoods of locally varying cell states. Variance of a feature is analyzed over each defined neighborhood and compared with the global variance of that feature, with a threshold of selection for downstream analysis. Selected features exhibit small local variance with gradual and monotonic changes, consistent with progressively transitioning cell states. In addition, Furchtgott et al. developed a Bayesian approach for identifying subsets of

gene expression patterns over 3 cell states that are useful for defining lineage relationships [64]. These feature selection methods use unique patterns of gene expression present in single-cell data sets to filter out genes whose variances are either owing to noise or are irrelevant to the phenomenon of interest. More refined gene expression patterns perhaps can be identified in the future for more sophisticated feature selection.

However, these algorithmic approaches perform best when provided with high-quality features, which are often confounded by artifacts such as gene dropouts, resulting from incomplete transcriptomic sampling, and stochasticity, arising from the amplification of single-cell scale reaction materials [65]. Detectable cell-to-cell variation can also originate from stochastic gene expression, where an underlying level of randomness is captured at the time of sample processing [66,67]. These sources of variation necessitate machine learning strategies for the selection of biologically meaningful features [68].

Unsupervised feature selection algorithms such as neighborhood variance ratio (NVR), dpFeature (dpF), FindVariableGenes (FVG) and PCA-Based Feature Extraction (PCAFE) are distinct strategies for achieving this goal in the context of pseudotemporal analysis [62,63,69,70]. Although feature selection is essential, the assumptions and performance of these algorithms have not been systematically evaluated, confounding the applicability of these methods to different datasets. Work by Chen et al. examined an underlying characteristic of high-dimensional data that interacts with these algorithms with a focus on NVR and dpFeature [71].

This algorithm presented by Welch et al. generates a connected graph based on the Euclidean distances of cell-to-cell gene expression. Based on this graph, the algorithm compares the variance of gene expression within neighborhoods and the variance of gene expression globally on a cell-to-cell basis. It then assumes that if the neighborhood variance is lower than the global

variance, there exists some meaningful and controlled gene expression. The formalization of this neighborhood variance, in the context of genes, is described as follows, where n is the sample number, kc is the minimum number of neighbors in the connected graph, g is the gene of interest, and N(i , j) is the nearest neighbor j of the sample i:

$$S_g^{2(N)} = \frac{1}{nk_c - 1} \sum_{i=1}^{n} \sum_{j=1}^{k_c} \left(e_{ig} - e_{N(i,j)g} \right)^2$$

Figure 10. Formal description of NVR algorithm.

An example of this phenomenon would be the expression of some gene that changes monotonically along the progression of a given developmental lineage. Neighborhood variation would be low given the gradual change of gene expression, and global variation would be higher given the differences in expression between end states of a transition. Due to the calculation of neighborhood variance, the time complexity of this algorithm is O(n) where n is the product of the number of cells and the number of genes. The following is the pseudocode for the algorithm:

1. Determine the minimum number of connections, k, that will generate a connected graph.
    a. Calculate the pairwise distances between each element of the input matrix
    b. Convert this vector into squareform
    c. Generate an adjacency matrix based on this squareform
    d. Permit k number of connections and generate a graph based on the adjacency matrix
        i. Count the number of connected components, c
        ii. If C>1, add 1 to k and repeat until C=1
2. Use this number of connections, k, to generate a connected graph

3. For each gene, calculate the mean variance of some n neighbors based on the generated graph
    a. Repeat for all possible neighborhoods
    b. Calculate the mean of this neighborhood variance
4. For each gene, calculate the global variance in the context of all cells
5. If the global variance of a gene divided by the average neighborhood variance of that same gene is greater than 1, select that gene.

A popular feature selection method, developed by the Trapnell group [62], utilizes density peak clustering [72] on a t-SNE (t-distributed stochastic neighbor embedding) dimension-reduced representation of transcriptomic data [73]. For t-SNE, our study used the monocle R package using the parameters max_components=2, num_dim=6, and check_duplicates=FALSE. Using this representation of the data, density peak clustering was performed. A generalized linear model was then used to test for the most significantly differentially expressed genes between clusters.

Principal component analysis (PCA)-based unsupervised feature extraction (FE) is a another method used to select biologically relevant genes [70]. This method starts by scaling the raw count data and performing a principal component analysis. For the first three principal components, the gene weights are then scaled and summed. These sums are used for a Chi- squared test. Finally, an adjusted p-value threshold is set and genes that meet that threshold are selected.

Following these explorations into feature selection and data pre-processing, we devised an open source pipeline incoporating these methods **(Figure 11)**. This pipeline consists of bioinformatic read alignment with the STAR aligner, droplet count matrix estimation with DropEst, and preliminary quality control with the scRNABatchQC R package [74–76]. First, dropTag (part of DropEst) takes paired-end, raw .fastq files and tags them in the context of unique molecular

identifiers (UMIs) and cellular barcodes for the demultiplexing process. This is dependent on the scRNA-seq platform's barcode whitelist; in this case we use the inDrop V1 and V2 barcodes. Before running the actual alignment process, a genome index must first be generated with respect to the reference and annotation files. STAR is a fast, scalable RNA-seq aligner which has splice awareness and takes the multiple tagged fastq.gz files generated by dropTag and aligns them using a reference genome index. The sorted .bam file generated by STAR alignment is used as an input to dropEst, which generates a barcode by gene count matrix, or droplet matrix, from the STAR aligned transcripts. Finally, scRNABatchQC is used to provide summary statistics and a quality assessment of the generated droplet matrix. This droplet matrix is further filtered in the heuristic droplet filtering and automated droplet filtering with dropkick section variants.

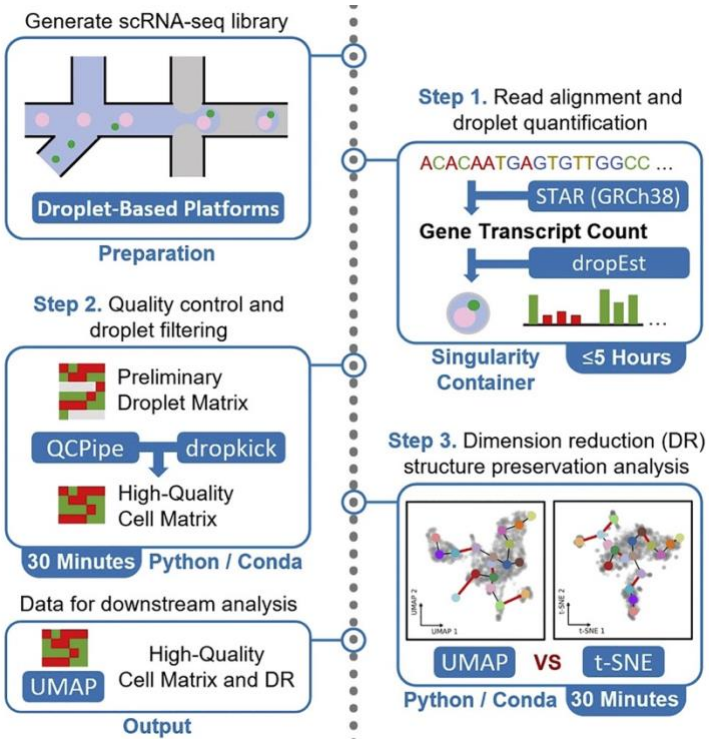


Figure 11. Open-source pipeline developed for the processing of scRNA-seq data.

Our pipeline is a three phase protocol which can be modularly tuned for various widely used scRNA-seq platforms.

Once these data exist in cell matrix form, barcode filtering is necessary, and can be used modularly if the user has a pre-computed matrix, either from the single-cell read alignment and DropEst library quantification section of this protocol or an external source, so long as the rows represent cell barcodes and columns represent genes. The output for this section is a cell matrix, differing from a droplet matrix in that it only contains gene read counts from only high-quality, intact single cells. First, a data-driven cutoff, by means of finding the inflection point in a cumulative sum curve of ranked barcode counts, is generated and used to minimize information-spars barcodes. This cumulative sum curve method borrows from an electrical engineering, through the usage of operating points. Like the determination of a time interval threshold in human-computer interactions or a stable operating point of an AC motor, we used a data-driven heuristic for the detection of high-quality libraries with this cutoff optimization. This method is related to what Satopää et al. refers to as “knee point” finding, and is a heuristic often used to find operating points in complex systems. Formally, this “knee” is akin to geometric curvature previously described for any continuous function f as:

$$K_f(x) = \frac{f''(x)}{\left(1+f'(x)^2\right)^{1.5}}$$

Figure 12. Formal description of Kneedle algorithm.

Where Kf (x) represents a standard closed-form defining the curvature of f at any point as a function of its first and second derivative. We then devised an algorithm which maximized the curvature for some threshold (cell quality, time thresholds, rotational speed, etc.), x.

After this first pass matrix filter, another distribution of uniquely detected genes per droplet is automatically thresholded through Otsu’s method, separating the remaining information-rich and

information-sparse droplets and generating a binary metadata label. Third, tissue-specific gene expression signatures are visualized after DR to pinpoint cell populations of interest for downstream analysis. Fourth, unsupervised clustering is performed to and evaluated to consistently discretize the single-cell transcriptional landscape [77]. Finally, by heuristically integrating these metrics and expression signatures, populations of intact single-cells and their respective high-quality transcriptomes can be saved for downstream analyses.

Multi-omic Mappings of Cellular Developmental Trajectories

A disruption to hub nodes of biological interactions often leads to the dysregulation of otherwise homeostatic equilibria, which vary across human tissues composed of heterogeneous systems of cells. This dysregulation may alter developmental processes that result in disease states capable of propagating throughout multicellular systems, ultimately affecting as macroscale symptoms, diagnoses, and prognoses. Pinpointing the disrupted junctions of developmental processes remains a high-value target for computational biology. These junctions may exist as transient cell states, as defined through the epigenome, transcriptome, or proteome, and contribute to the dynamic nature of tissue heterogeneity in multicellular organisms. Accordingly, multi-omic snapshots of this heterogeneity may capture cell states across a range of developmental lineages, which can be rationally ordered in high-dimensional feature-spaces to summarize a continuum of cellular differentiation [78,79]. By examining these relationships, developmental timelines and "pseudo-timelines" can be mapped out under the assumption that such multicellular snapshots capture self-renewing and differentiating communities of cells. With omics-level data at high enough resolutions, relative comparisons between homeostatic conditions and disease states can be made, where cell state intersections represent potential points of meaningful dysregulation. Three major classes of algorithms are popularly used to map out these multicellular states, being trajectory inference, clonal inference, and RNA velocity.

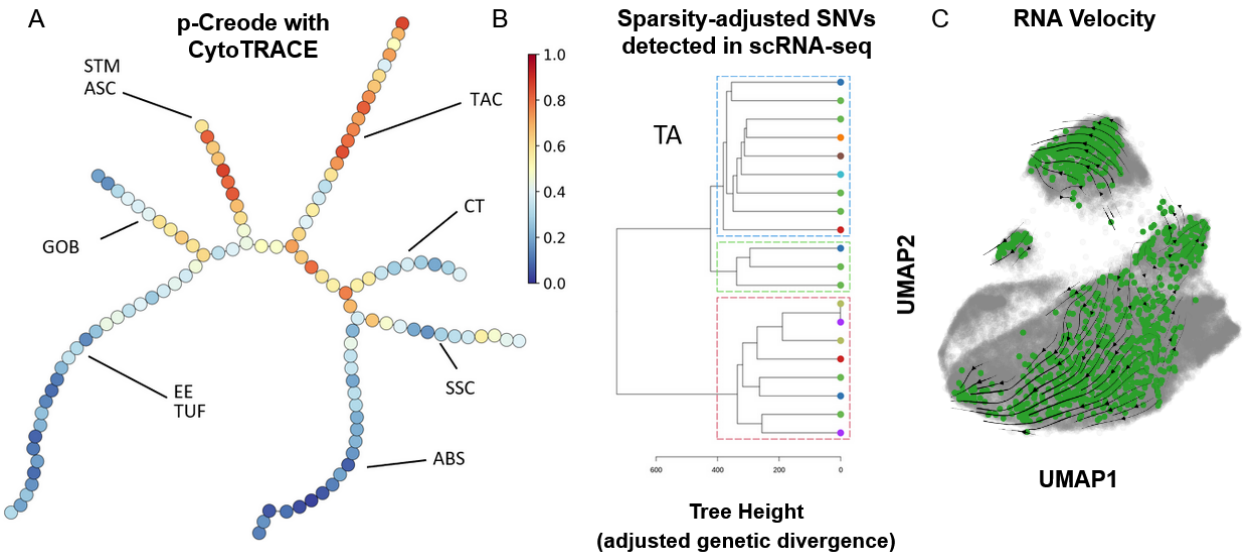


Figure 13. Computational methods for inferring developmental dynamics.

(A) p-Creode trajectory of colonic epithelium and tumor tissues overlaid with CytoTRACE scores (B) Modified DENDRO phylogenetic estimations for the same libraries as in (A) (C) RNA velocity of normal cells highlighted on UMAP of libraries in (A).

Trajectory inference methods are typically implemented in R or Python with wrappers around functions written in lower-level programming languages and generally take single-cell count matrices as input. pCreode is an algorithm which uses an ensemble of density-dependent k-Nearest Neighbors (k-NN) in principal component space. Using this ensemble of graphs, the most representative graph structure was chosen by minimizing the Gromov-Hausdorff distance [68]. The result is a hierarchical tree of transcriptional variation across developmental pseudotime points, where branchings represent splits of developmental processes into different lineages. Similarly, Monocle 3 employs principal graph learning through PAGA (partition-based graph abstraction) [62]. In three steps, Monocle 3 reduces the dimension of the count matrices, partitions the data into supergroups (consisting of communities of cells detected through the Louvain algorithm) using the PAGA algorithm, and finally learns a smoothed principal graph using these supergroups as

landmark points. The resulting principal graph, or multidimensional graph, is embedded into three dimensions using SimplePPT for interpretability. Given the appropriate transformations, even epigenomic single-cell data, representing snapshots of dynamic cell states can be mapped out. Namely, the STREAM (Single-cell Trajectories Reconstruction, Exploration And Mapping) algorithm can be generalized from scRNA-seq to scATAC-seq data through the mapping of an elastic principal graph [80].

Clonal inference methods, unlike trajectory inference, utilize information retrieved from nucleotide variation and genotypes detected within each single-cell sequencing library. These are valuable since a core assumption of trajectory inference is that cells with similar transcription or epigenetic states are also developmentally similar, which may not be universally true as gene expression can vary greatly throughout the life cycle of a cell. In this case, the direct measurement of genetic history is used to reconstruct developmental lineages as opposed to an inferred genetic history based on gene expression or regulation. The molecular content of genomes, however, remains relatively difficult to recover intact from a single-cell. While some methods have used pseudo-bulk aggregates of cells instead, others have applied rationally-devised, statistical adjustments to transcriptomic data to detect germline genetic variation. Calling genetic variation in transcribed sequences still remains a challenge, given that an additional layer of variation in dynamic gene expression is also captured. Examples of this would include high or low expressing genes which differ between two distinct subpopulations of cells, where the quantification of genotypic reads is confounded by an inherent increase in the likelihood of detecting highly expressed transcripts. DENDRO or DNA based EvolutionNary tree preDiction by scRNA-seq technOlogy, uses a Beta binomial generative model to describe these stochastic effects of gene expression by considering gene expression in the description of genotypes from transcriptomic sequencing data [81]. scLineager, similar to DENDRO, employs a Bayesian hierarchical model. This model takes factors

and their hierarchical relationships into account such as gene dropout or stochastic gene expression alongside their correlations to known cell line genetic variation [82].

RNA velocity methods, like clonal inference, also make of information that exists in the sequence of transcripts. Instead of dissecting the phylogenetic relationships, the splice states of detected reads are instead used to describe potential developmental states. Differing from both trajectory inference and clonal inference, RNA velocity predicts the likelihood of cell state transitions and their directionality *de novo*. In conjunction with splice-aware sequence alignment for single-cell libraries, the velocyto python package was introduced with the concept of RNA velocity in seminal work done by Le Manno et al. [83] Subsequent improvements in the prediction of cell state vector fields have been through the incorporation of dynamical modeling, as described by Bergen et al. and implemented in the scVelo python package [84].

Transcriptomically-Derived Gene Signatures and Regulatory Networks

Although experimental and computational methods can now be used reliably to measure the whole transcriptomes of single cells, the abstraction of biological function from these transcriptomes remains a challenge. On one hand, top-down approaches reliant on published literature are highly dependent on existing bodies of knowledge and are largely used for the summarization of pre-defined gene expression programs. Summarization is important, as genes seldom act in a vacuum; the joint evaluation of multiple genes comprising individual biological programs better describes biologically meaningful phenomena as opposed to the quantification of single gene features in isolation. Gene signature scoring, like that described by Satija et al., provides a robust method to calculate the enrichment of gene sets compared to a randomly generated background signature [85]. Such methods come with an increased risk of committing type II errors, given that biological processes uncharacterized by existing literature could not be directly measured nor summarized. On the other hand, purely bottom-up approaches have a

higher risk of committing type I errors, given that co-varying biological processes may be functionally disjoint in practice. cNMF, or consensus Non-negative Matrix Factorization, robustly decomposes an observation-by-feature matrix into two component matrices, denoted H and W, which when multiplied together reconstructs the original gene expression matrix [86]. This is typically a computationally intensive process which explores the space of possible H and W combinations, and in the case of cNMF, is instead an ensemble of sampled factorizations. In this case H and W can be reframed as matrices representing gene expression programs and program representation per cell. Pragmatically, the utilization of both top-down and bottom-up methods, such as with the SCENIC algorithm occupy a space which allows for a mixture of mining regulatory networks *de novo*, while also being able to incorporate prior knowledge about transcription factor protein-DNA interactions, compiled from literature [87]. Similar to cNMF, SCENIC detects gene co-expression modules *de novo*, using gradient boosting instead of NMF. To parse potential false positives, by virtue of coincidental co-expression, the SCENIC pipeline uses known transcription factor regulatory motifs to assess whether co-expressed genes share a common co-expressed transcription factor. Importantly, this step draws from the biological intuition that gene expression is tightly controlled in regulatory networks, largely under the regulation of transcription factors specifically interacting with the genome.

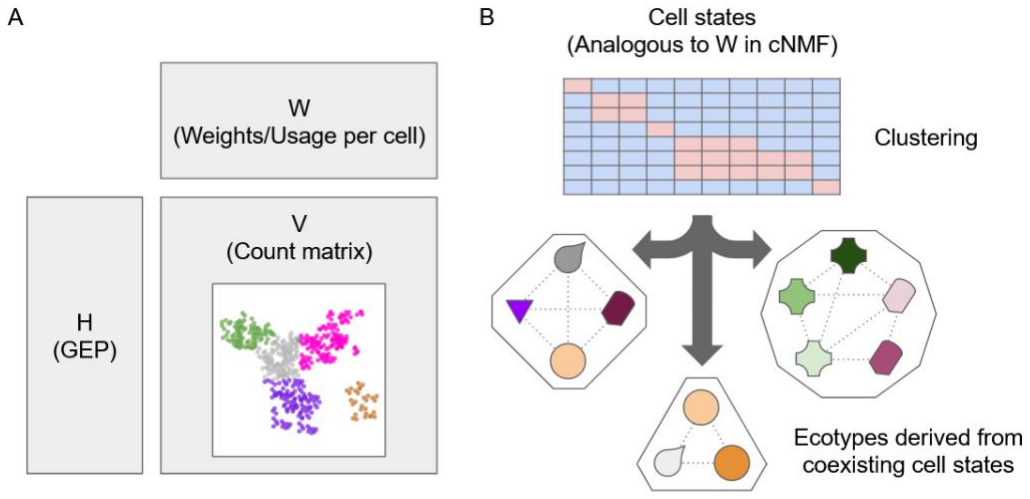


Figure 14. Gene expression program inference based on matrix factorization-based.

(A) cNMF decomposition of matrix V results in W and H, which represent “usages” or weightings per cell and gene expression programs respectively. When multiplied, V can be reconstructed. (B) Ecotyper detects coexisting cell states by clustering a decomposed matrix, similar to the W of cNMF.

Inferring multicellular interactions through scRNA-seq

While the calculation of gene signature or regulatory network enrichment can abstract biologically meaningful, cell-intrinsic insights from single-cell transcriptomes, meaningful interactions between those inferred processes require additional sets of analysis. Modeling the gestalt of a system of cells should include the consideration of models with higher orders of complexity in the form of cell-extrinsic regulatory networks alongside cell-intrinsic models [88,89]. Like cell-intrinsic network inference, an analogous dichotomy between top-down and bottom-up is also seen in its algorithmic strategies. Instead of coexpressed genes within a single-cell or cell-type, the abstraction of cell-extrinsic biological processes is dependent on the paired expression of receptor-ligand pairs and their respective mediators. Top-down methods, again, are dependent on the retrieval of intercellular receptor-ligand pairs mined from literature and large-scale databases. Popular algorithms such as Cellphonedb and iTalk are dependent on combinatorial statistical testing for the discovery of significantly paired receptor-ligand gene expression [90,91]. From these gene sets, queries are made to manually curated databases of receptor-ligand subunits, finally outputting lists of interactions alongside their probabilistic matrices. Bottom-up approaches follow a similar trend to their cell-intrinsic counterparts. Primarily, these methods employ versions of matrix factorization algorithms, not unlike cNMF, in finding correlated gene expression programs across distinct populations of cells. Namely, Ecotyper and multicellular immune hub prediction [92,93]**.** These methods are heavily dependent on the non-negative factorization of gene expression across multiple, co-captured cell populations of heterogeneous

tissues. Gene expression programs decomposed from these factorizations are then scanned for co-regulation across these populations, leading to pathway-level inferences of cell-cell interactions. Alternatively, NicheNet blends top-down and bottom-up processing, like SCENIC [94]. This method employs prior knowledge to refine data-driven pathway definitions, which may provide a lower likelihood of type I or II errors, though few meta-analyses have thoroughly evaluated the performance of these methods given their recency.

Features of the Colonic Epithelium

The molecular characterization of heterogeneous tissues has more recently been used for the understanding of the mammalian colon, a complex system of neural, connective, endothelial, epithelial, and immune cells of further specification [95]. Compounding these interactions, the lumen of the gastrointestinal tract contains diverse microbial populations comprising the gut microbiome, and its interactions with the host organism's complex systems have yet to be deeply characterized. The macro-structural organization of the colon can be classified by its anatomical positioning as well as its cross-section layers [96]. Anatomically, the colon or large intestine is attached to the ileum of the small intestine, or the most distal segment. In proximal-to-distal order, the colon can be generally divided into five segments: the cecum and appendix, ascending colon, transverse colon, descending colon, and sigmoid colon. Importantly, the luminal properties of the colon are heterogeneous given its proximal-distal axis [97]. Three examples include the speed of distal colonic transit (monotonic decrease), microbial diversity in the lumen (monotonic increase), and the volume of fecal short-chain fatty acids (monotonic decrease).

The cross-sectional organization of the colon involves four major layers of cells of varying thickness. In external-to-internal order, these layers are the serosa, muscularis, submucosa, and the mucosa [95]. The serosa, largely consisting of connective tissue, provides the outermost layer of the colon interfacing with the mesentery, which contains blood vessels and components of the

enteric nervous system. The muscularis contains the outermost layer of circular and longitudinal muscles, and with its innervation, this layer is important for initiating peristalsis. The submucosa is a second layer of connective tissue, which, like serosa and muscularis, contain specialized subsets of tissue including lymphatic vessels. Finally, the innermost layer of the colon is the mucosa, which contains a high diversity of self-renewing epithelial cell types, intermixed with a variety of non-epithelial stromal cells. This mucosa is the primary interface with the lumen of the gut, regularly enduring the vast majority of microbial insults [98,99].

A large focus on gastrointestinal biology remains on the micro-scale, molecular characterization of this luminal epithelium. Several pathogenic mechanisms may originate from this layer likely due to its rate of self-renewal, the complexity of interactions involving the immune system, and its proximity to the diversity of microbes and chemical signals [100]. Similar to the small intestine, intestinal glands, or crypts of Lieberkühn, line the inner surface of the mucosa, leading to an increase in surface area for the potential absorption of nutrients. Differing from the small intestine, colonic glands only contain a crypt domain, without a villus domain extending past the opening of the crypt and into the lumen. In terms of cellular organization and crypt compartmentalization, a gradient of morphgenic and juxtacrine signals maintain a diversity of undifferentiated, self-renewing stem cells and its differentiated descendants in homeostatic conditions [101–105]. Morphogenic signals such as those used by the BMP, Wnt, and EGF pathways involve the ligand-based transference of information across a distance and are dependent on the concentration of these cell-free ligands. Juxtacrine, or contact-dependent, signals such as Ephrin and Notch signaling, are reliant on the direct interaction of membrane-bound or extracellular-matrix originating ligands with membrane-bound receptors. The movement of molecules through a shared cellular junction, like a gap junction, would also enable this type of signaling. Other aspects affecting this signaling gradient include the concentration of oxygen, given that endothelial vessels reside closer to the base of the crypt, which has significant effects on hypoxia-inducible factor

signaling and its related effects on metabolism. In each of these cases, the gradient concentration of signals along the crypt leads to a series of coordinated pathway regulation which compartmentalize cellular differentiation and function.

Several molecularly distinct cell types reside at the base of the crypt, where Wnt, EGF, Notch, and Ephrin signaling is high [101–106]. Generally, these basally enriched cell types are related to the function or maintenance of stem cells, or multipotent and undifferentiated epithelial cells. These stem cells can generally be identified by their capacity to self-renew with symmetrical division, with functional studies yielding specific gene markers related to the regulation of these processes such as LGR5, LRIG1, AXIN2, and OLFM4 [107–111]. The maintenance and regulation of these progenitor cells is performed by differentiated deep-crypt secretory cells, which intercalate the stem cells at the crypt base [112,113]. With molecular function resembling the Paneth cells of the small intestine, these cells help to maintain the gradient of signaling molecules required for proper stem cell regenerative function and differentiation. Separated by a label retaining cell at the +4 position of the crypt, the transit amplifying cells act as a buffer between undifferentiated and differentiated cells [111,114]. These transit amplifying cells (TAC), originating from the adjacent stem cells, undergo rapid proliferation, and their daughter cells are physically displaced across the gradient of signals which, ultimately, leads to the differentiation into more mature cell types [101,115]. TACs, thus, can easily be identified by both their positioning and expression of genes involved in cell cycle phase progression, such as MKI67 and PCNA.

Differentiation downstream from stem cells and TACs lead to two primary epithelial lineages with more specific molecular functions, being the absorptive and secretory lineages. The absorptive lineage exists as a gradient of enterocytes at various points of developmental maturity. Crypt-top cells represent a fully mature state which typically reside at the tops of each crypt, expressing the genes BEST4 and OTOP2 which contribute to the maintenance of luminal pH [116–118]. Importantly,

these cells help to regulate homeostatic conditions interacting with luminal microbes. Colonocytes comprise the vast majority of cells in the colonic crypt and intercryptal surfaces. They exist in a gradient of differentiation along the axis of the crypt, and primarily act as mediators of nutrient and water absorption. Expectedly, both of these cell types are the most exposed to the genotoxic, microbial interactions, being located closest to the lumen and the opening of the crypt. The secretory lineage differs from the absorptive lineage with the expression of the transcription factor, ATOH1, an early cell fate determinant which inhibits the expression of HES1 [68,119]. Functionally, secretory lineage cells, of which there are three major subtypes, generally involve the secretion ranging from mucins to hormones, and are far outnumbered by absorptive lineage cells. The most common of these subtypes are the goblet cells, morphologically resembling a goblet spilling over in liquid, secrete high volumes of heavily glycosylated molecules known as mucins through the production of mucin-filled vesicles [96]. The resulting matrix of mucins, produced through genes such as MUC2, acts as a protective buffer between the lumen and epithelium. Other secretory-lineage cells are rarer in nature and have highly-specific functions, namely the enteroendocrine and tuft cells [68,120]. Enteroendocrine cells help to regulate the interaction between the gut and the endocrine system through the production of hormonal peptides secreted through pathways involving CHGA and CHGB [121]. Finally, tuft cells, named for their apical tufted protrusions, act as chemosensory immune mediators which are marked by the expression of DCLK1 and play a role in the detection of parasitic infections [122,123].

This epithelial complexity further interacts with non-epithelial systems of varying lineages, which include cells that are hematopoietic, mesenchymal, or endothelial in origin. Such cells generally reside within the stromal microenvironment, interacting with the wider immune system and acting as an interface between the mucosal epithelium and often modulated by disease states [124]. Hematopoietic cells produce two primary sub-lineages of immune cells that can be found interacting with the colonic epithelium, being lymphoid and myeloid cells [125]. Lymphoid cells

constitute a significant portion of the adaptive immune system in this microenvironment, can be further subdivided into T, B, and Natural Killer (NK) cells. T cells originate in a naïve state in the thymus and gain antigen specificity through interacting with antigen-presenting cells. Upon activation, these cells differentiate into two subtypes marked by the expression of CD8 or CD4. Similar to Natural Killer cells, CD8+ T cells gain cytotoxic function through the expression of perforins and granzymes to compromise the cellular membranes of its targets. Contrasting these, CD4+ cells are further specialized for the regulation of this cytotoxicity through the production of cytokines and chemokines. NK cells, unlike these CD8+ and CD4+ subtypes, maintain their cytotoxicity without a stepwise activation through the adaptive immune system. B and plasma B cells, while lymphoid in origin, differ greatly in function from the T cells. B cells and their descendents, such as plasma B cells, function as part of the adaptive immune system in presenting antigens and produce antibodies, which bind to the surface of its targets and allow for more specific immune interactions.

Paired with these lymphoid lineage cells, myeloid cells within the colonic microenvironment are represented primarily by granulocytes, macrophages, and dendritic cells [125]. Granulocytes, named for the appearance of their cytoplasmic granules, encompass several subclasses of myeloid cells which specialize in producing enzymes, such as lysozymes, which help to break down pathogenic agents. The granulocyte class includes mast cells and neutrophils, for example. Mast cells in the gut often interact with the enteric nervous system and act as regulators of inflammation, by secreting proinflammatory cytokines. Neutrophils are also part of the innate immune system and release enzymes through degranulation, while also undergoing phagocytosis in the neutralization of pathogens. Monocytic immune cells, as part of the myeloid lineage, include macrophages and dendritic cells, both of which are antigen presenting cells. Macrophages, like neutrophils, phagocytose pathogenic materials, process, and present the relevant antigens for utilization by the adaptive immune system. Dendritic cells share a similar role at the interface

between the innate and adaptive immune systems. Structurally, the dendrites extending from these cells increase their effective surface areas for microenvironmental sampling and secretion of signaling molecules such as inflammatory cytokines.

Cell-intrinsic and extrinsic characterizations of colorectal cancer

Colorectal cancer (CRC) is the fourth most common cancer and second leading cause of cancer death in the United States [126]. Classification schemes for CRC focus largely on intrinsic features of tumor cells, including histopathology, bulk gene expression (Consensus Molecular Subtypes or CMS), chromosomal instability (CIN), hypermethylation (CpG Island Methylator Phenotype or CIMP), and microsatellite-instability (MSI) [127–129]. Additionally, the tumor microenvironment and immune response is critical to CRC pathogenesis, as highlighted recently [93]. Hypermutated MSI-high (MSI-H) tumors exhibit a neoantigen-triggered cytotoxic immune infiltration that contributes to their responsiveness to immunotherapy [124,130]. However, a significant subset of low mutation burden CRCs appears to exhibit an activated immune microenvironment via ill-defined mechanisms [131]. We hypothesize that mapping the routes towards tumorigenesis in precursors of MSI-H and MSS CRCs will uncover mechanisms that define the CRC cellular landscape and identify targets with diagnostic or therapeutic utility.

Most MSS and MSI-H CRCs develop from pre-cancerous conventional adenomas (ADs) and sessile serrated lesions (SSLs; formerly sessile serrated adenomas/polyps), respectively. As proposed by Vogelstein and others, ADs arise from truncating mutations in APC, which result in activation of the WNT pathway and CIN [132]. ADs subsequently accumulate gain-of-function mutations in oncogenes (chiefly KRAS) and loss-of-function mutations in tumor suppressor genes such as TP53, ultimately forming MSS CRCs. Conversely, SSLs resemble MSI-H CRCs molecularly and are distinct from ADs in that tumorigenesis is not initiated by genetic disruptions of APC [133,134]. Instead, they have epigenetic disruptions, including MLH1 hypermethylation and a

40-75% prevalence of CIMP [135,136]. These tumors harbor BRAF mutations in contrast to KRAS mutations commonly present in ADs. Mirroring the relatively lower incidence of MSI-H CRCs and their prevalence in the proximal colon, SSLs represent only 10-20% of polyps and are also found more often in the proximal colon, unlike the more frequently distal ADs [133,137,138].

Chapter III - Gene regulatory networks and the discovery of polyp-specific transcriptional programs from heterogeneous tumor tissues

Recreated from:

Chen, B., Scurrah, C. R., Mckinley, E. T., Simmons, A. J., Ramirez-Solano, M. A., Zhu, X., Markham, N. O., Heiser, C. N., Vega, P. N., ... Coffey, R., Shrubsole, M., & Lau, K. S. (2021). Differential pre-malignant programs and microenvironment chart distinct paths to malignancy in human colorectal polyps. Cell. https://doi.org/10.1016/j.cell.2021.11.031

Introduction

We present a multi-omic human pre-cancer atlas integrating single-cell transcriptomics, genomics, and immunohistopathology describing the two most common pathways towards CRC. We identify and functionally validate distinct origins and molecular processes that establish divergent tumor landscapes. Notably, this clearer understanding of advanced and highly heterogeneous cancers was enabled only by looking at CRCs through the lens of their originating lesions, paving a path to new strategies for precision prevention, surveillance, and therapeutics.

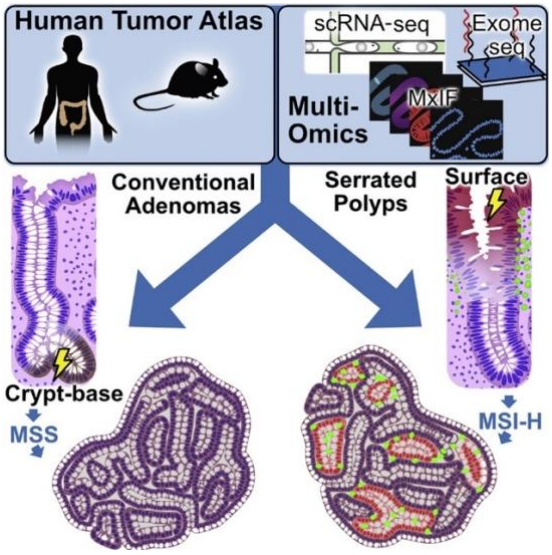


Figure 15. Overview of findings from pre-cancer atlas.

This human tumor atlas was derived from multi-omic human data and validation in mouse and organoid experiments. Shown are the two divergent routes of tumorigenesis, dependent on cell position within the crypt, and ultimately resulting in malignancies with contrasting microenvironmental properties.

Distinct histopathologic and molecular features define colonic pre-cancer subtypes

Polyps, as well as matching normal biopsies, were collected from COLON MAP study participants recruited as described in the methods. Most polyps were small (median diameter ≤ 5 mm) and were bisected for multi-assay analysis. Single-cell RNA-seq (scRNA-seq), multiplex immunofluorescence (MxIF), and multiplex immunohistochemistry (MxIHC) were performed on two independent sets of specimens collected approximately 1 year apart. The Discovery (DIS) set consisted of 65 specimens analyzed including 30 tumors. The Validation (VAL) set consisted of 63 specimens analyzed including 32 tumors **(Figure 16A)**. Overall, 128 independent scRNA-seq datasets on 62 tumors were generated **(Figure 16A)**. Specimens were collected from diverse sex, racial, and age groups. In addition, we performed bulk RNA-seq and targeted gene sequencing on an orthogonal set of 66 and 281 polyps, respectively **(Figure 17A)**.

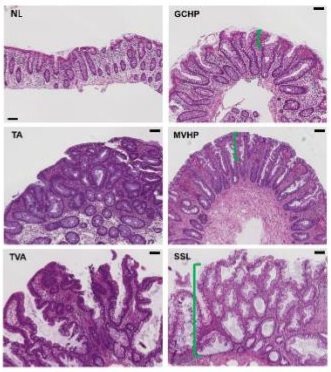
A

Pre-cancer Polyp Sample Sets				Colorectal Cancer Sample Sets	
TCPS	COLON MAP Discovery (DIS)		COLON MAP Validation (VAL)	VUMC	Broad
Polyps	Polyps	Normals	Polyps	Normals	Cancers	Cancers
301	27	35	28	31	33	60

Assay	Conventional Adenomas (AD)	Serrated Polyps (SER)	Normal Biopsies (NL)	Microsatellite Stable (MSS)	Microsatellite Instability-High (MSI)
scRNA-seq	29	19	66	33	34
Whole Exome-seq	17	16	0	17	15
MxIF or MxIHC	20	23	0	17	15

RNA-seq	58	8
Targeted DNA-seq	275	6

B



C

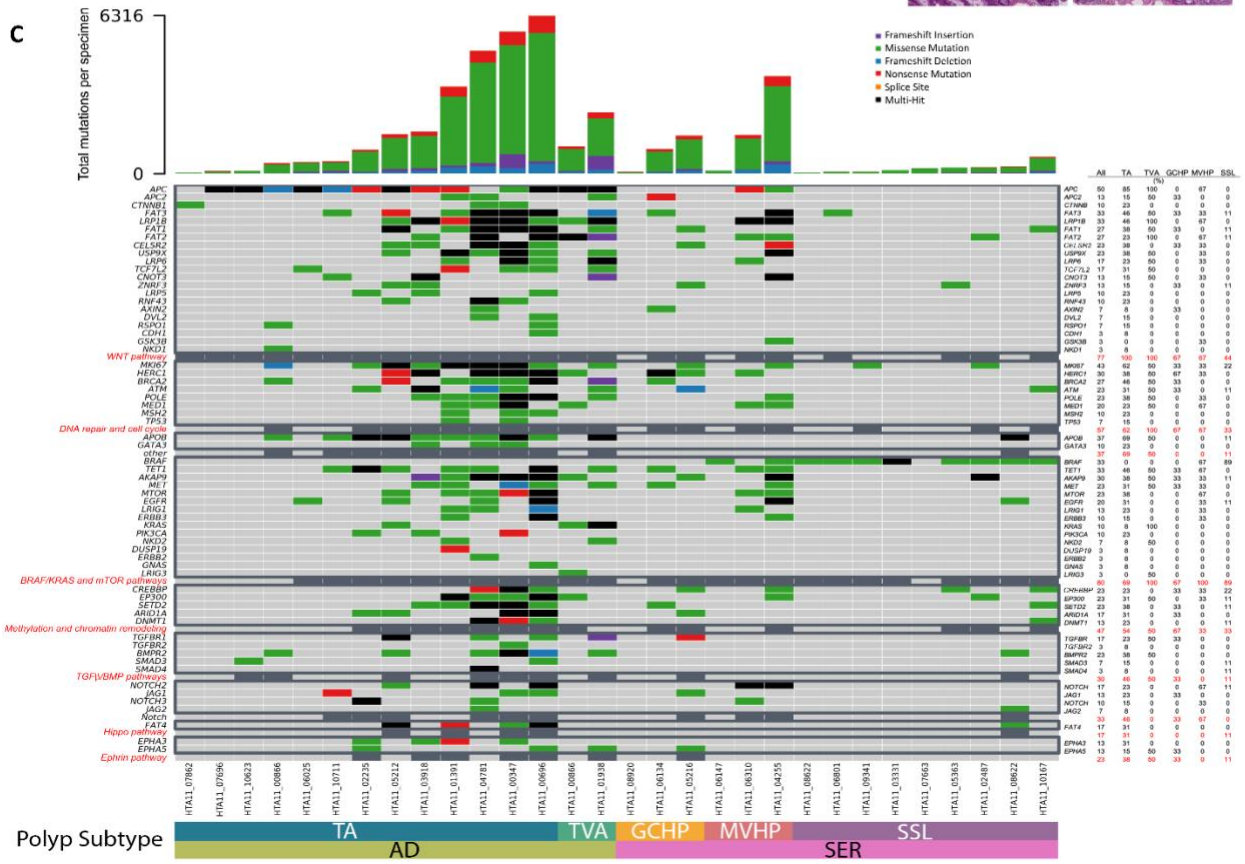
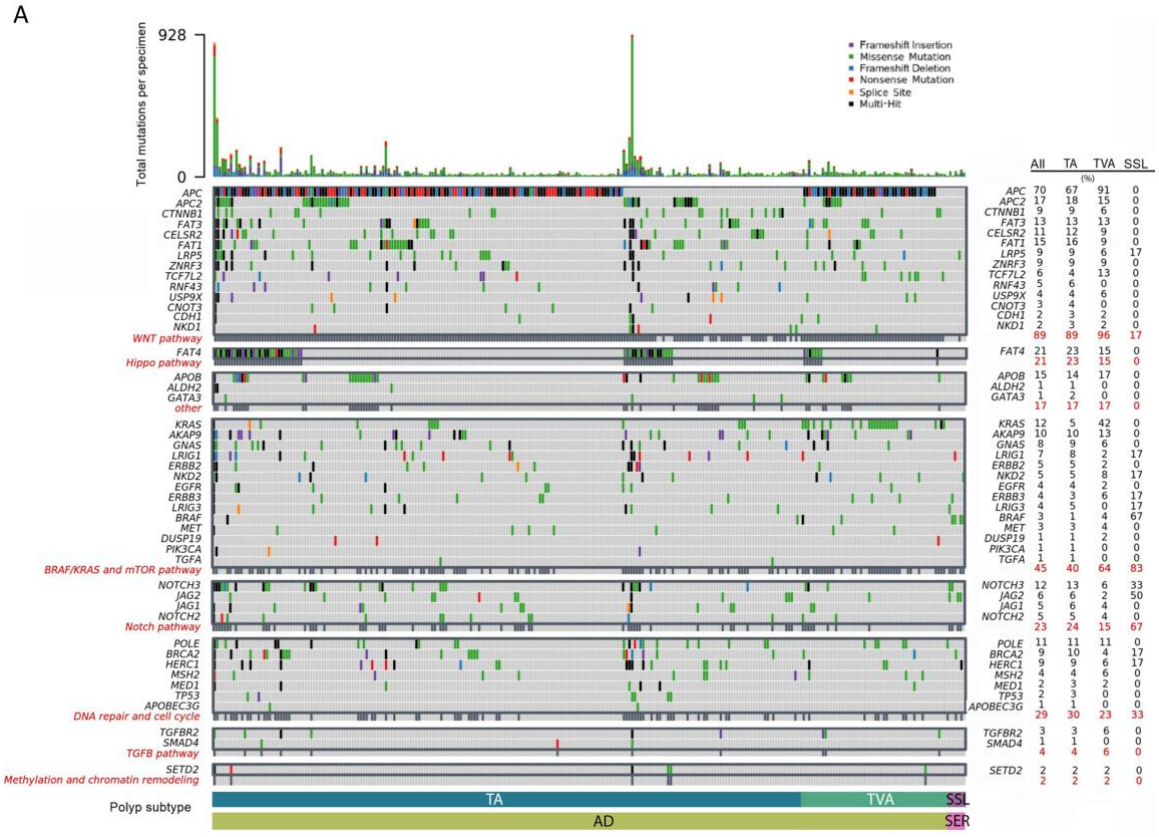


Figure 16. Histological and mutational features of human colonic pre-cancers.

(A) Experimental design for the multi-omic characterization of pre-cancers and CRCs, with subtypes classified based on histopathology, exome-sequencing based mutational spectra, and MSI-testing. Two independent datasets were collected for each group of sample sets: the DIS and VAL datasets for colonic biopsies and polyps and the VUMC and Broad datasets for CRCs. **(B)** Representative H&E (Haematoxylin and Eosin) images depicting the histology of normal colonic tissue and polyp subtypes in the study. Green brackets label portions of crypts occupied by neoplastic cells. **(C)** Oncoplot representation of the mutational landscape of 30 polyps analyzed by exome sequencing and somatic mutation calling. Total numbers of mutations detected per specimen represented by a bar plot (top), and different types of mutations color-coded. Important genes for CRC are grouped into pathways (dark gray boxes). Percentage of gene and pathway mutations within polyp subtypes summarized in a table (right). Multi-hit refers to multiple mutations in the same gene detected within a specimen.


A
928
0
Total mutations per specimen
Frameshift Insertion
Missense Mutation
Frameshift Deletion
Nonsense Mutation
Splice Site
Multi-Hit
All TA TVA SSL
(%)
APC 70 67 91 0
APC2 17 18 15 0
CTNNB1 9 9 6 0
FAT3 13 13 13 0
CELSR2 11 12 9 0
FAT1 15 16 9 0
LRP5 9 9 6 17
ZNRF3 9 9 9 0
TCF7L2 6 4 13 0
RNF43 5 6 0 0
USP9X 4 4 6 0
CNOT3 3 4 0 0
CDH1 2 3 2 0
NKD1 2 3 2 0
WNT pathway 89 89 96 17
FAT4 21 23 15 0
Hippo pathway 21 23 15 0
APOB 15 14 17 0
ALDH2 1 1 0 0
GATA3 1 2 0 0
other 17 17 17 0
KRAS 12 5 42 0
AKAP9 10 10 13 0
GNAS 8 9 6 0
LRIG1 7 8 2 17
ERBB2 5 5 2 0
NKD2 5 5 8 17
EGFR 4 4 2 0
ERBB3 4 3 6 17
LRIG3 4 5 0 17
BRAF 3 1 4 67
MET 3 3 4 0
DUSP19 1 1 2 0
PIK3CA 1 1 0 0
TGFA 1 1 0 0
BRAF/KRAS and mTOR pathway 45 40 64 83
NOTCH3 12 13 6 33
JAG2 6 6 2 50
JAG1 5 6 4 0
NOTCH2 5 5 4 0
Notch pathway 23 24 15 67
POLE 11 11 11 0
BRCA2 9 10 4 17
HERC1 9 9 6 17
MSH2 4 4 6 0
MED1 2 3 2 0
TP53 2 3 0 0
APOBEC3G 1 1 0 0
DNA repair and cell cycle 29 30 23 33
TGFBR2 3 3 6 0
SMAD4 1 1 0 0
TGFB pathway 4 4 6 0
SETD2 2 2 2 0
Methylation and chromatin remodeling 2 2 2 0
Polyp subtype
TA
TVA
SSL
AD
SER

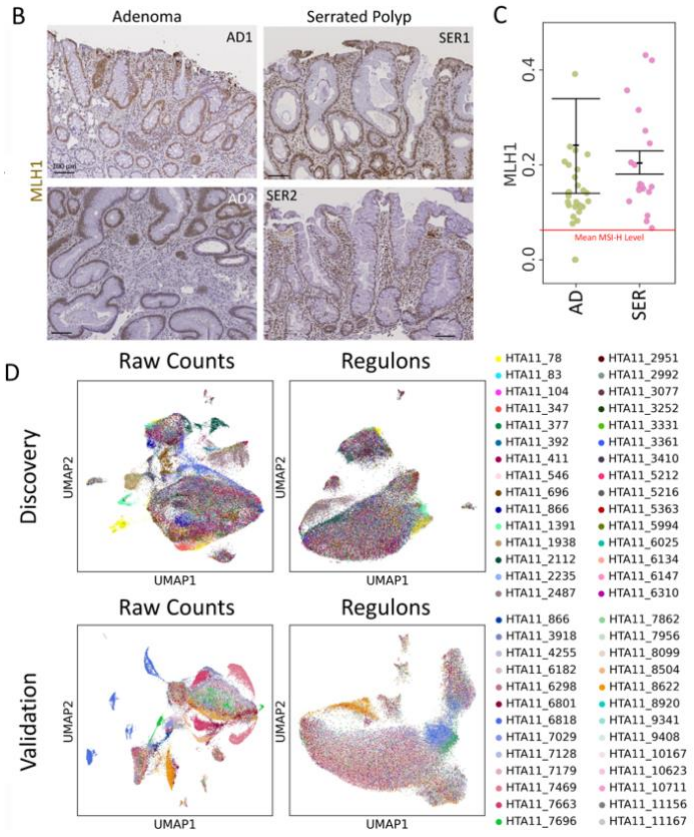

B
Adenoma
Serrated Polyp
AD1
SER1
AD2
SER2
MLH1
C
MLH1
0.4
0.2
0.0
Mean MSI-H Level
AD
SER
D
Raw Counts
Regulons
Discovery
Validation
UMAP1
UMAP2
HTA11_78
HTA11_83
HTA11_104
HTA11_347
HTA11_377
HTA11_392
HTA11_411
HTA11_546
HTA11_696
HTA11_866
HTA11_1391
HTA11_1938
HTA11_2112
HTA11_2235
HTA11_2487
HTA11_2951
HTA11_2992
HTA11_3077
HTA11_3252
HTA11_3331
HTA11_3361
HTA11_3410
HTA11_5212
HTA11_5216
HTA11_5363
HTA11_5994
HTA11_6025
HTA11_6134
HTA11_6147
HTA11_6310
HTA11_866
HTA11_3918
HTA11_4255
HTA11_6182
HTA11_6298
HTA11_6801
HTA11_6818
HTA11_7029
HTA11_7128
HTA11_7179
HTA11_7469
HTA11_7663
HTA11_7696
HTA11_7862
HTA11_7956
HTA11_8099
HTA11_8504
HTA11_8622
HTA11_8920
HTA11_9341
HTA11_9408
HTA11_10167
HTA11_10623
HTA11_10711
HTA11_11156
HTA11_11167

Figure 17. Heterogeneous molecular landscapes of human colonic pre-cancers.

(A) Oncoplot representation of the mutational landscape of 281 polyp specimens detected through bulk targeted sequencing and mutational variant calling, with layout similar to Figure 1C. **(B)** Representative IHC for MLH1 in AD and SER. **(C)** Quantification of *MLH1* gene expression in AD and SER from scRNA-seq of polyp-specific cells, as compared to the same analysis of MSI-H CRC cells (red line - mean). n=29 for AD and 19 for SER. Each point represents an individual, averaged sample using normalized and scaled Arcsinh-transformed single-cell counts. **(D)** UMAP representation of epithelial scRNA-seq data, color coded by specimen to depict interspecimen heterogeneity, examining (top) DIS and (bottom) VAL sets, and (left) raw count-based and (right) SCENIC regulon-based.

Pre-cancerous polyps were histologically categorized by two pathologists into two subtypes: ADs consisting of tubular ADs (TAs) and tubulovillous ADs (TVAs), or serrated polyps (SERs) consisting of hyperplastic polyps (HPs) and SSLs. TAs were much more common than TVAs in both the DIS and VAL sets, and they were found on both sides of the colon. SSLs were found preferentially on the proximal side of the colon consistent with their expected distribution. Representative polyps are shown in **Figure 16B**. TAs had less than a 25% villous component, whereas the more histologically advanced TVAs had between 25-75% villous features. True villous adenomas (> 75% villous features) were not found in either set. TAs and TVAs exhibited varying degrees of conventional dysplasia with characteristic elongated, pseudostratified nuclei and increased mitotic activity **(Figure 16B)**. HPs and SSLs were categorized based on morphology at the crypt base and distribution of epithelial serrations. HPs are subdivided into goblet cell-rich HPs (GCHPs) and microvesicular HPs (MVHPs). Although the malignant potential of HPs is debated [135], MVHPs appear to be relatively more advanced and may progress to SSLs [133]. GCHPs had enlarged crypts with numerous goblet cells throughout the crypt length; epithelial serrations, if present, were subtle and confined to the mucosal surface **(Figure 16B)**. MVHPs had elongated crypts, many of which contained microvesicular mucin granules, and fewer goblet cells at the crypt base. Epithelial serrations extended from the surface to two thirds down the crypt, sparing the crypt base. For both types of HPs, the crypt base was morphologically normal. In contrast, SSLs showed epithelial serrations that extended to the base of crypts, which were dilated and spread laterally above the muscularis mucosae; goblet cells were also found throughout the crypt **(Figure 16B).** SERs infrequently displayed overt cytologic dysplasia.

Next, we characterized the mutational profiles of ADs and SERs by conducting exome sequencing and detecting somatic mutations by comparison to paired blood or buccal cell specimens to remove germline sequence variants **(Figure 16C)**. Due to small polyp sizes and the prioritization of fresh tissue for single-cell assays, we used the clinical Formalin Fixed Paraffin Embedded

(FFPE) material from both sample sets for exome sequencing, and about half generated sufficient sequence quality for further analysis. Although the numbers of samples were low, the predominant patterns of mutations were consistent with published literature. *APC* mutations were detected in 86% (11/13) of the TAs and in both TVAs. Only one (8%) TA had a *KRAS* mutation, while both TVAs harbored *KRAS* mutations, consistent with TVAs being more histologically advanced. Mutations in *FAT1-4*, multifunctional genes involved in WNT and Hippo pathways, were observed frequently in ADs. All but one of the 9 SSLs (89%) had the oncogenic *BRAFV600E* mutation; none of the 3 GCHPs harbored *BRAF* mutations but 2 (67%) MVHPs did, consistent with MVHPs being SSLs in evolution. Neither *APC* nor *KRAS* mutations were detected in any of the SSLs, and none of the ADs had *BRAF* mutations. Somewhat surprisingly, none of the SSLs exhibited a hypermutation phenotype, while a portion of TAs/TVAs did. Whereas MLH1 expression is usually lost in MSI-H CRCs due to promoter methylation, MLH1 protein and gene expression in SSLs were comparable to ADs, both of which were higher than the mean MSI-H CRC level **(Figure 17B,C).** Biallelic loss in mismatch repair genes was not detected in any polyp, further supporting that these SSLs had not yet acquired a hypermutation phenotype.

We validated this mutational analysis using targeted gene sequencing of a separate larger set of 281 premalignant tumors, consisting mostly of TAs and TVAs **(Figure 17A)**. Once again, *APC* mutations were found in 67% (148/222) of TAs and this increased to 91% (48/53) in TVAs. Likewise, mutations in *KRAS* increased markedly from TAs (5%, 12/222) to TVAs (42%, 22/53). *BRAF* mutations were again enriched in SSLs (67%) compared with TAs (1%) and TVAs (4%). Again, none of the SSLs exhibited a high mutation load, where several TA/TVAs did, confirming exome sequencing results. Only a single mutation in the TGF-b/BMP pathway was observed in SSLs. Non-*APC* mutations in WNT pathway genes, such as *RNF43* or *ZNRF43*, were not common in SSLs from either dataset. When signaling pathways were queried from the combined mutational analysis, a picture emerges of WNT-driven tumorigenesis in TA and TVAs, but SSLs

are less dependent on this pathway. Most, but not all SSLs harbor *BRAF* mutations; yet, they are not hypermutated.

Single-cell analysis identifies neoplastic cells in conventional adenomas and serrated polyps that arose from distinct tumorigenic processes

We generated scRNA-seq data on 70,691 (DIS studies) and 71,374 cells (VAL studies), (Total: 142,065), respectively, after filtering for high quality barcodes using dropkick [139]. For pre-cancer-specific analysis, we selected a subset of 121 (DIS:62 and VAL:59) scRNA-seq datasets across normal biopsies and all four pre-cancer polyp subtypes as defined by histological classification. Cells from specimens with unconfirmed histology (labeled UNC) were transcriptomically classified. Overall, 55 total polyps were analyzed **(Figure 16A)**. We conducted UMAP dimension reduction on raw scRNA-seq data and observed intermixing of epithelial cells from normal colonic biopsies and immune cells from different participants, indicating the absence of batch effects **(Figure 17D)**. However, neoplastic tissues clustered by sample, demonstrating intertumoral variability consistent with unique tumorigenic processes.

Since transcription factor (TF)-defined regulon activities are considered to be determinant of cell identity in a transcriptomic landscape, we used SCENIC (Single-Cell rEgulatory Network Inference and Clustering), which is a regulon-based, batch-robust feature extraction tool, to adjust for polyp-specific effects [87,140]. This process factors out environmentally sensitive and/or random contributors such as metabolic genes. Clustering and co-embedding 62 samples from the DIS dataset in regulon space, we identified nine major epithelial cell populations **(Figure 18A,B)**. Biopsies from normal colonic tissues served as reference landmarks for seven canonical epithelial cell types, including goblet cells (*MUC2/ATOH1+)*, absorptive cells (*KRT20/GUCA2A+*), crypt top colonocytes (*BEST4/MEIS1+*), enteroendocrine cells (*CHGA/NEUROD1+)*, tuft cells (*POU2F3/SOX9+)*, transit-amplifying cells (*PCNA/MKI67+)*, and stem cells (*LGR5/OLFM4+)*

(Figure 18A,B; Figure 19A). Polyp samples also contained substantial numbers of normal cells consistent with the histopathology **(Figure 16B, 18B; Figure 19B)**. However, two cell populations were overwhelmingly represented in polyp samples, as determined by the sample-by-sample breakdown of proportional cluster representation **(Figure 18B,C; Figure 19C)**. One population was enriched in TA and TVA, hereafter referred to as ASCs (AD-specific cells, p<1E-4 MWU test). The second neoplastic population was enriched in SSLs and HPs, hereafter referred to as SSCs (serrated-specific cells, p<1E-4 MWU test). Importantly, these results, as well as others below, were consistent across DIS and VAL datasets **(Figure 18A-C; Figure 19A-D)**, which demonstrate rigor and reproducibility of our data and enable high-confidence identification of ASCs and SSCs for further analysis.

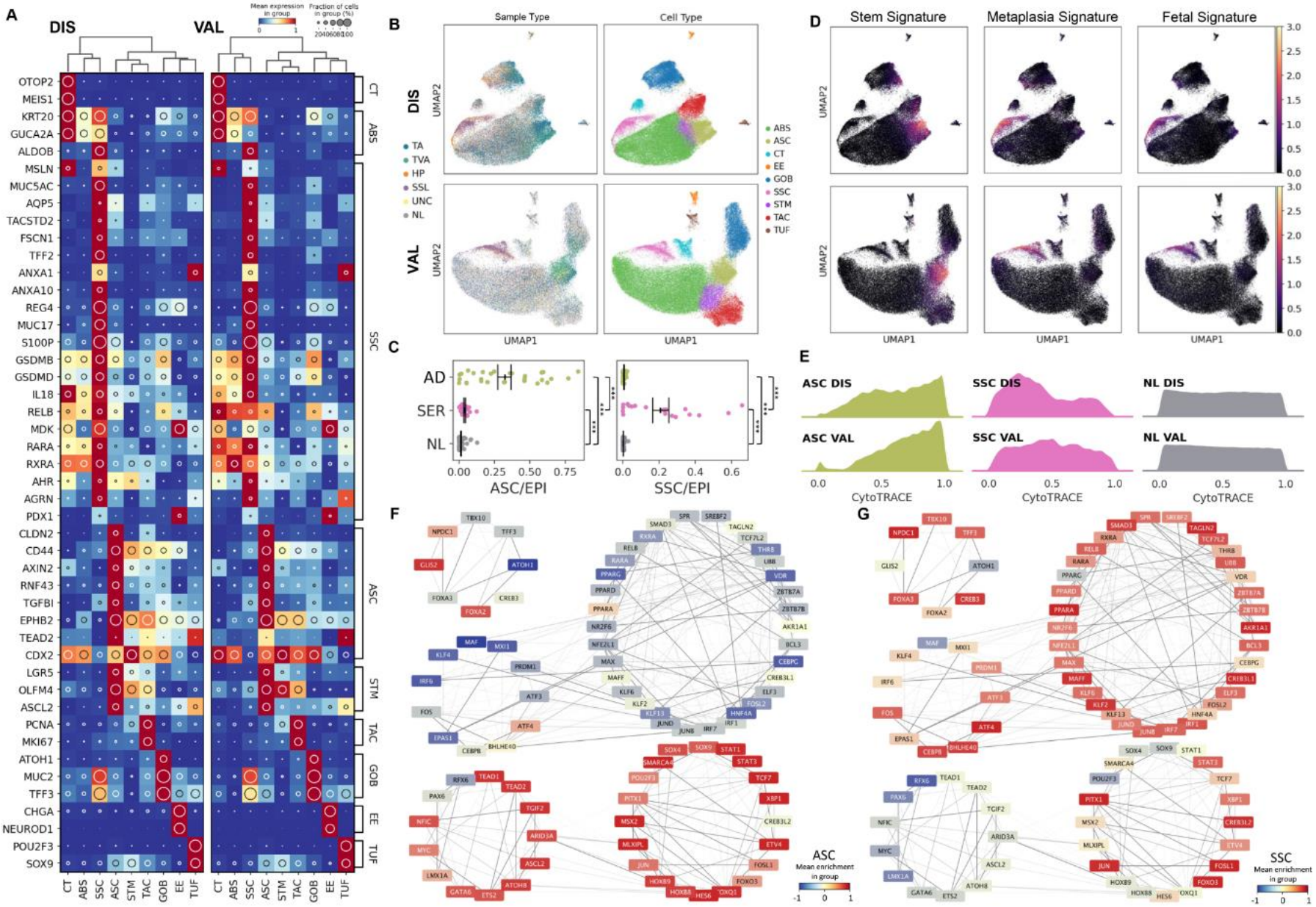

A
DIS
VAL
Mean expression in group
Fraction of cells in group (%)
OTOP2
MEIS1
KRT20
GUCA2A
ALDOB
MSLN
MUC5AC
AQP5
TACSTD2
FSCN1
TFF2
ANXA1
ANXA10
REG4
MUC17
S100P
GSDMB
GSDMD
IL18
RELB
MDK
RARA
RXRA
AHR
AGRN
PDX1
CLDN2
CD44
AXIN2
RNF43
TGFBI
EPHB2
TEAD2
CDX2
LGR5
OLFM4
ASCL2
PCNA
MKI67
ATOH1
MUC2
TFF3
CHGA
NEUROD1
POU2F3
SOX9
CT
ABS
SSC
ASC
STM
TAC
GOB
EE
TUF
B
Sample Type
Cell Type
DIS
VAL
UMAP2
UMAP1
TA
TVA
HP
SSL
UNC
NL
ABS
ASC
CT
EE
GOB
SSC
STM
TAC
TUF
C
AD
SER
NL
ASC/EPI
SSC/EPI
D
Stem Signature
Metaplasia Signature
Fetal Signature
E
ASC DIS
SSC DIS
NL DIS
ASC VAL
SSC VAL
NL VAL
CytoTRACE
F
G
ASC
SSC
Mean enrichment in group

Figure 18. Single-cell gene expression and regulatory network landscape of conventional and serrated polyps.

(A) Heatmap representation of top biologically relevant and differentially expressed genes for (left) DIS (n = 62) and (right) VAL (n = 59) datasets. Inset circle indicates prevalence within defined single-cell populations and color intensity represents scaled and standardized Arcsinh gene expression. Cell types as defined by Leiden clustering and marker gene detection. ABS-absorptive cells, ASC-adenoma-specific cells, CT-crypt top colonocytes, EE-enteroendocrine cells, GOB-goblet cells, STM-stem cells, SSC-serrated-specific cells, TAC-transit amplifying cells, TUF-tuft cells. **(B)** Regulon-based UMAP of (top) DIS and (bottom) VAL datasets color overlaid with (left) tissue and polyp subtype and (right) cell type. **(C)** Scatter plots of normalized (left) ASC or (right) SSC representation per tissue and polyp subtype. Points represent individual specimens. Error bars represent SEM of n = 29 for AD, n = 19 for SER, and n = 66 for NL. **(D)** Stem, metaplasia, and fetal signature scores overlaid onto UMAP of (top) DIS and (bottom) VAL datasets. **(E)** Ridge plots of CytoTRACE score distributions for ASC, SSC, and NL cell populations across (top) DIS and (bottom) VAL datasets. **(F,G)** TF target network created from normal and pre-cancer cells, organized into super-regulons derived from shared targets for **(F)** ASCs and **(G)** SSCs. Color overlays for each TF node are averaged and normalized regulon enrichment scores, while edge opacities are the inferred TF-target weightings. *p<0.05, **p<0.01, ***p<0.001.

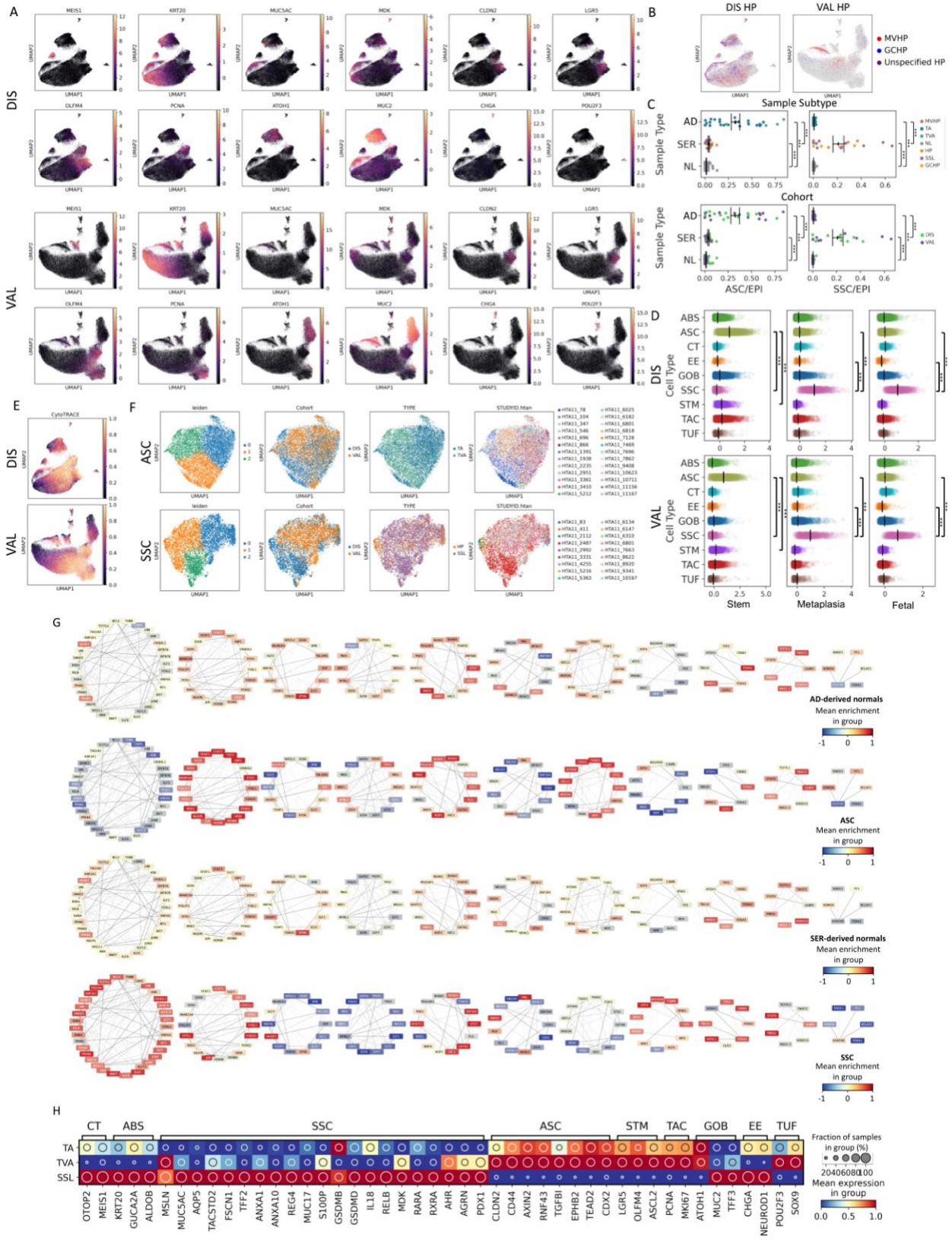

A
DIS
VAL
MEIS1
KRT20
MUC5AC
MDK
CLDN2
LGR5
OLFM4
PCNA
ATOH1
MUC2
CHGA
POU2F3
UMAP1
UMAP2
B
DIS HP
VAL HP
MVHP
GCHP
Unspecified HP
C
Sample Subtype
Cohort
Sample Type
AD
SER
NL
MVHP
TA
TVA
NL
HP
SSL
GCHP
DIS
VAL
ASC/EPI
SSC/EPI
D
DIS
VAL
Cell Type
ABS
ASC
CT
EE
GOB
SSC
STM
TAC
TUF
Stem
Metaplasia
Fetal
E
CytoTRACE
F
ASC
SSC
leiden
Cohort
TYPE
STUDYID.htan
G
AD-derived normals
ASC
SER-derived normals
SSC
Mean enrichment in group
H
CT
ABS
SSC
ASC
STM
TAC
GOB
EE
TUF
TA
TVA
SSL
OTOP2
MEIS1
KRT20
GUCA2A
ALDOB
MSLN
MUC5AC
AQP5
TACSTD2
FSCN1
TFF2
ANXA1
ANXA10
REG4
MUC17
S100P
GSDMB
GSDMD
IL18
RELB
MDK
RARA
RXRA
AHR
AGRN
PDX1
CLDN2
CD44
AXIN2
RNF43
TGFBI
EPHB2
TEAD2
CDX2
LGR5
OLFM4
ASCL2
PCNA
MKI67
ATOH1
MUC2
TFF3
CHGA
NEUROD1
POU2F3
SOX9
Fraction of samples in group (%)
Mean expression in group

Figure 19. Heterogeneity of colonic polyps depicted by scRNA-seq data.

(A) UMAP of regulon-based epithelial cell scRNA-seq data overlaid with cell type-specific genes. Color intensity represents scaled and standardized Arcsinh gene expression. **(B)** UMAP of (left) DIS and (right) VAL datasets color overlaid with HP subtypes. **(C)** Scatter plots of normalized (left) ASC or (right) SSC representation per tissue and polyp subtype. Points represent individual specimens. Error bars represent SEM of n = 29 for AD, n = 19 for SER, and n = 66 for NL. Coloring of points indicates (top) polyp subtype or (bottom) DIS or VAL dataset. **(D)** Scatter plots of signature scores by cell type, with each point representing a single cell, for (top) DIS and (bottom) VAL datasets. Error bars depict SEM of single cells. Not all post-hoc tests shown. **(E)** Regulon-based UMAP of epithelial cell scRNA-seq data overlaid with CytoTRACE scores for (top) DIS and (bottom) VAL datasets. **(F)** Regulon-based UMAPs of (top) ASC and (bottom) SSC-gated scRNA-seq data with colored label overlays indicating Leiden subcluster, cohort, polyp subtype, and specimen ID. **(G)** Expanded TF target network created from normal and pre-cancer cells, organized into super-regulons derived from shared targets for AD-derived normals, ASCs, SER-derived normals, and SSCs. Color overlays for each TF node are averaged and normalized regulon enrichment scores, while edge opacities are the inferred TF-target weightings. **(H)** Heatmap representation of gene sets derived from the scRNA-seq applied to bulk RNA-seq data of colonic polyps (n=36 tubular, 22 tubulovillous, 8 SSLs). Inset circle indicates the fraction of specimens presented with gene expression and color intensity represents mean scaled and standardized Arcsinh gene expression. *p<0.05, **p<0.01, ***p<0.001.

Next, we identified gene programs and pathways differentially activated in ASCs and SSCs compared to accompanying normal epithelial cells. ASCs resembled colonic stem and progenitor cells and expressed genes indicative of WNT pathway activation (*LGR5*, *OLFM4*, *ASCL2*, *AXIN2*, *RNF43*, and *EPHB2*) **(Figure 18A)**. The stem cell signature in ASCs was greater than crypt base stem cells from matching normal colon biopsies from the same individuals **(Figure 18D; Figure 19D)** [106,107,109,110,141–143]. Crypt base-specific genes such as *CLDN2* and *CD44* were also enriched in ASCs, suggesting ADs originate from the bottom of colonic crypts. Because ASCs resembled normal stem cells, we used CytoTRACE to infer their stem potential [144]. First, we demonstrated normal colonic biopsies contained stem cells with high CytoTRACE scores transitioning into differentiated cells with lower CytoTRACE scores **(Figure 18E; Figure 19E)**. The score distribution was relatively uniform in normal biopsies due to a mix of stem, transitioning, and differentiated epithelial cells. In contrast, CytoTRACE analysis of ASCs yielded a distribution skewed towards cells with high predicted stem potential **(Figure 18E)**. This variation in stem potential suggests the presence of tumor stem cells, which is further supported by the relative enrichment of GO terms associated with WNT-driven stemness within specific ASC subclusters **(Figure 19F)**. Together, these analyses describe a model wherein WNT-dependent stem cell expansion initiates tumorigenesis in ADs, which is consistent with known WNT pathway-activating gene mutations prevalent in CRC initiation, most notably loss-of-function mutations in *APC* [145].

In marked contrast to ASCs, SSCs did not exhibit WNT pathway activation or a stem cell signature, but instead shared transcriptomic similarities with differentiated cells (**Figure 18A,D; Figure 19A,D**). The CytoTRACE scores of SSCs were skewed towards lower predicted stem potential, opposite to ASCs **(Figure 18E; Figure 19E)**. Given low overall stemness, heterogeneous populations of SSCs with variable differentiation characteristics were still observed **(Figure 19F)**. The transcriptomic profiles of SSCs resembled absorptive-lineage cells, but SSCs also expressed functional goblet cell genes, including *TFF3* and *MUC2*. Unlike normal

goblet cells, they did not express the master secretory cell TF *ATOH1*, and the *ATOH1*-related regulon was not enriched, suggesting SSCs harbor a mixed cellular identity **(Figure 18A; Figure 19G)**. To this point, SSCs highly expressed genes not normally observed in the colon (*MUC5AC*, *AQP5*, *TACSTD2* (TROP2), *TFF2*, *MUC17, and MSLN*), but rather found in other endodermal organs. Most notably, SSCs expressed a gastric metaplasia gene signature not expressed in ASCs or normal colonic cells **(Figure 18A,D; Figure 19D)** [146–151]. This surprising finding, along with the expression of differentiated cell gene signatures in SSC, led us to hypothesize metaplasia may underlie the pathogenesis of SSLs.

Metaplasia is a process by which differentiated cells transdifferentiate to non-native cell types, often occurring as a regenerative mechanism after damage. Loss of *CDX2*, a hindgut homeobox TF, in the colon is associated with an imperfect pyloric-type gastric metaplasia and a shift towards expression of genes more rostral in the rostral-caudal gradient [152–154]. In our datasets, *CDX2* was expressed in most colonic cell types, including ASCs; however, it was downregulated in SSCs, supporting a loss of regional identity in these cells **(Figure 18A)**. This loss of caudal identity in SSCs was accompanied by a reversion to an embryonic stage as evident by a fetal gene expression signature; this includes the *MDK* gene, which encodes a heparin-binding growth factor only transiently expressed early in normal colonic development **(Figure 18A,D; Figure 19A,D)** [155]. Luminal communication pathways function through receptors for retinoic acid (RA) (*RXRA/RARA*) and the aryl hydrocarbon receptor (*AHR*), and these were enriched in SSCs **(Figure 18A)** [156–159]. A recent paper reported luminal RA aided in the maturation of absorptive cells while suppressing a YAP-dependent regenerative stem state in organoids [160]. Absorptive genes stimulated by RA, such as *ALDOB*, were similarly increased in SSCs. However, rostral identity genes suppressed in absorptive cell differentiation, such as *ANXA10* (gastric) and *ANXA1* (fetal and esophageal), were upregulated in SSCs [147,151,161,162] **(Figure 18A).** These gene

signatures depict a loss of colonic identity and provide further evidence SSCs arise from a metaplastic process.

In addition to differential gene expression, we used TF target similarity to create a common TF regulatory network depicting the coordinated regulation of genes as programs and pathways. Some coordinated clusters of regulons, which we referred to as super-regulons, were overrepresented in ASCs versus SSCs, including WNT-driven and Hippo-driven super-regulons marked by MYC, ASCL2, TCF7, and TEAD1 activities **(Figure 18F; Figure 19G)**. These results were consistent with the role of Hippo signaling and the ASCL2 transcriptional complex in regeneration and renewal responses of intestinal stem cells [163,164]. For SSCs, supporting the role of a damage-induced metaplastic process, a super-regulon indicating interleukin signaling and microbiota interaction was observed **(Figure 18G; Figure 19G).** Specifically, the upregulated transcription factor activities for SSCs included RELB (NF-κB signaling), IRF1, IRF6, and IRF7, reflecting the immunogenic state of these epithelial cells **(Figure 18A,G),** which was corroborated by gene set enrichment of microbial infection response, innate immune activation, and epithelial wound healing pathways [165]. Supporting the activation of interferon response elements, coordinated upregulation of inflammasome-related genes such as *IL18* and gasdermins further implicated responses to external pathogens as triggers of metaplasia **(Figure 18A)** [166–169]. Similarly, regulons related to FOSL2, KLF4, and ATF3 were enriched **(Figure 18G; Figure 19G)**, drawing parallels to recent work documenting increased chromatin accessibility of these TF targets in a mouse model of microbiota-driven colitis [170]. The increased presence of luminal communication regulons, driven by RXRA, RARA, and VDR, in SSCs further supports the potential involvement of luminal microbiota **(Figure 18G; Figure 19G)** [160,171,172]. The regenerative, immunogenic, and microbiota-responsive regulons in SSCs are all indicative of an active metaplastic program.

We performed bulk RNA-seq on an additional 58 ADs (36 TAs, 22 TVAs) and 8 SSLs to further validate our findings. The same gastric metaplasia signature in our single-cell data was enriched in the SSLs from this validation set (**Figure 19H**). We also observed the same paradoxical expression of goblet cell genes (*TFF3* and *MUC2*) without *ATOH1* in SSLs. The WNT-driven stem cell signature was present in ADs, with TVAs exhibiting slightly higher expression **(Figure 19H)**. These results validate our scRNA-seq findings of a WNT-activated program of stem cell expansion in ADs, and a program of gastric metaplasia, likely arising from a committed cell lineage, in SSLs.

Serrated polyps arise from a distinct cellular origin compared with conventional adenomas

Because SSCs may arise from metaplasia of differentiated cells, we hypothesized SERs originate from differentiated cells in a “top-down” model of tumorigenesis, compared to ADs arising from proliferative stem cells in a “bottom-up” fashion. To provide histological evidence of tumor origins and to build on our scRNA-seq results with spatial data, we mapped the location of neoplastic cells by multiplex histological and immunofluorescence imaging. Stem cell markers, OLFM4 and SOX9, were abundant in ADs but were significantly reduced in HPs and SSLs **(Figure 20A,B; Figure 21A,B)**. Nuclear CDX2 was detected in the normal colon and in ADs but was decreased in HPs and absent in SSLs **(Figure 20C; Figure 21C)**. MUC5AC, a marker of SSCs, was highly expressed in HPs and SSLs but was absent from normal colonic biopsies and ADs **(Figure 20D; Figure 21D)**. Interestingly, MUC5AC-positive, neoplastic cells were often observed at the top of the crypt with normal-appearing MUC5AC-negative cells at the crypt bottom, implying a non-crypt origin of SERs. MUC5AC-positive cells first appeared at the luminal surface in GCHPs and then extended further to the crypt base in MVHPs and SSLs **(Figure 21D,E)**, consistent with the histopathological progression of these SERs **(Figure 16B)** and supporting the luminal surface origin of SSCs. MUC5AC-positive cells were detected in the majority of abnormal crypts from SERs **(Figure 21F)**, indicating metaplasia is a homogeneous feature of these polyps. In the

normal colon, MUC5AC staining was largely absent, but occasional MUC5AC immunoreactivity was detected, again, at the luminal surface in a few specimens **(Figure 21G)**. MUC5AC staining was increased in the epithelial compartment of ulcerative colitis patients **(Figure 21G)**, supporting metaplasia as a response to epithelial damage. Luminal surface colonic cells appear susceptible to damage-induced metaplasia that may elicit serrated polyp formation if the damage is not resolved.

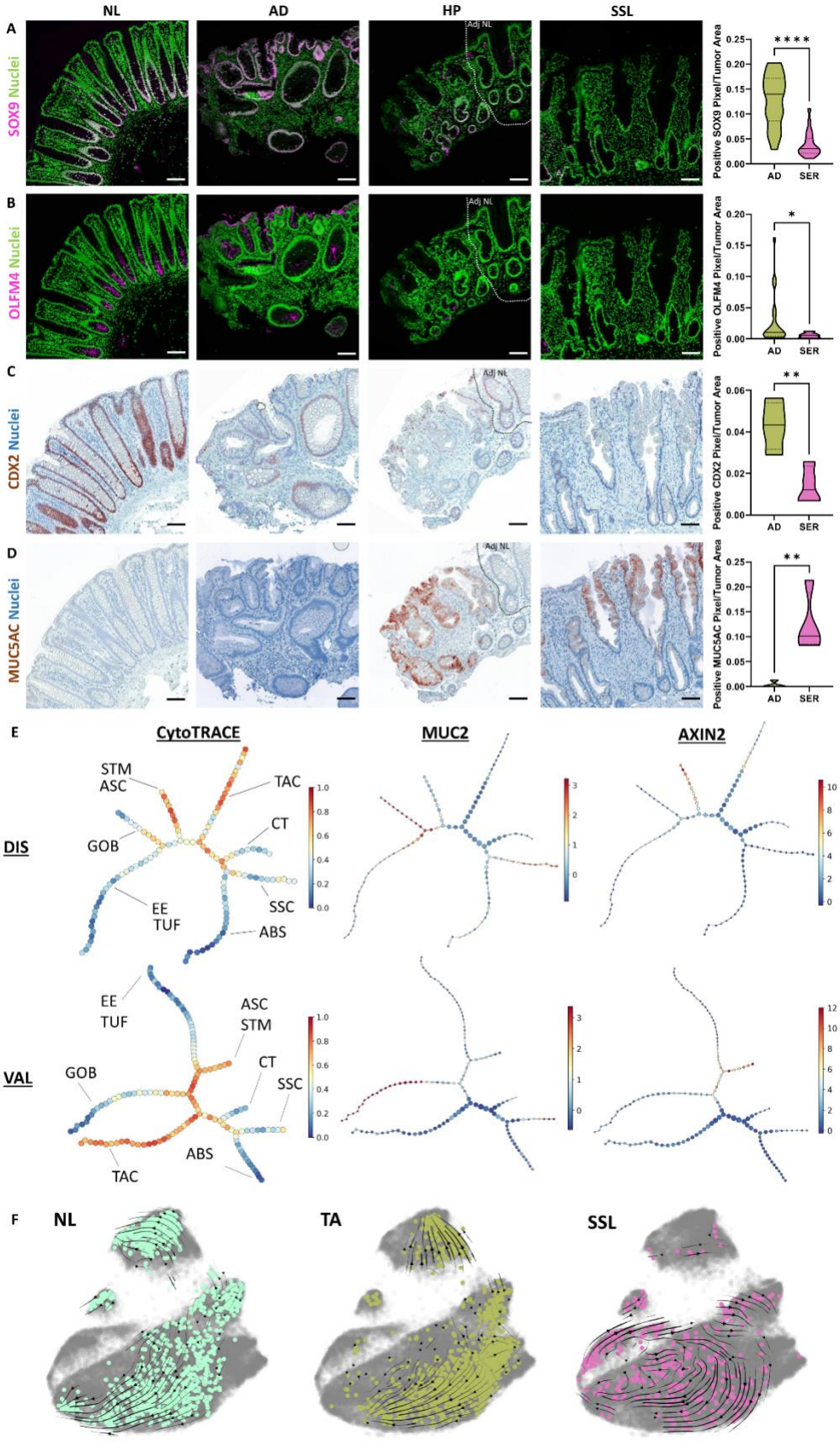

NL
AD
HP
SSL
A
SOX9 Nuclei
Adj NL
Positive SOX9 Pixel/Tumor Area
****
AD
SER
B
OLFM4 Nuclei
Positive OLFM4 Pixel/Tumor Area
*
C
CDX2 Nuclei
Positive CDX2 Pixel/Tumor Area
**
D
MUC5AC Nuclei
Positive MUC5AC Pixel/Tumor Area
E
CytoTRACE
MUC2
AXIN2
DIS
STM
ASC
TAC
CT
GOB
EE
TUF
SSC
ABS
VAL
F
NL
TA
SSL

Figure 20. Inferred origins of conventional and serrated polyps.

(A-D) Representative multiplex images of colonic polyps and normal tissues for **(A)** SOX9, **(B)** OLFM4, **(C)** CDX2, and **(D)** MUC5AC. Quantification of multiplex images (right) of positive pixels per tumor area for each marker of n=20 polyps per polyp type (AD vs. SER), presented as violin plots with median as solid lines and quartiles as dotted lines. **(E)** Trajectory inference through p-Creode performed on the regulon landscape of single epithelial cells. Overlay represents CytoTRACE scores. Individual branches representing developmental lineages labeled by canonical markers. Insets are p-Creode trajectories with *AXIN2* overlay to represent WNT pathway activity, and *MUC2* overlay to represent goblet cell mucin production. For these insets, node size represents the proportion of cells and overlay color intensity represents mean scaled and standardized Arcsinh gene expression. **(F)** RNA velocity for representative NL, TA, and SSL overlaid on combined UMAP embedding for DIS sample set. Vectors inferring average cell state transitions within each specimen are shown as black arrows. Colored points represent cells derived from the individual specimen, with grey points representing all other cells.

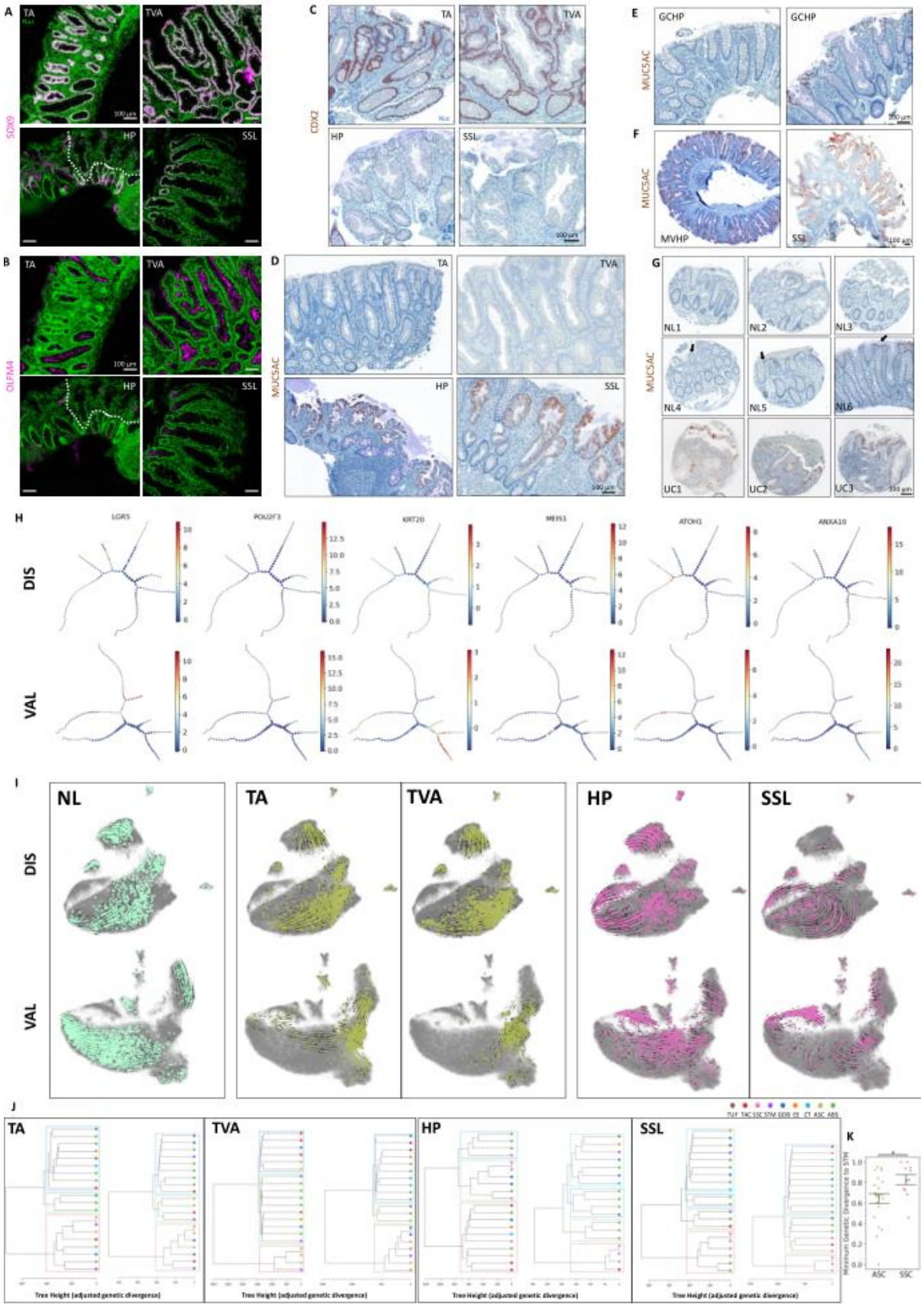

A
TA
TVA
HP
SSL
SOX9
B
OLFM4
C
CDX2
D
MUC5AC
E
GCHP
MUC5AC
F
MVHP
SSL
G
NL1
NL2
NL3
NL4
NL5
NL6
UC1
UC2
UC3
H
LGR5
POU2F3
KRT20
MEIS1
ATOH1
ANXA10
DIS
VAL
I
NL
TA
TVA
HP
SSL
DIS
VAL
J
TA
TVA
HP
SSL
Tree Height (adjusted genetic divergence)
K
ASC
SSC

Figure 21. Stem cell and metaplastic programs in conventional and serrated polyps.

(A-D) Additional representative multiplex images of colonic polyp subtypes for **(A)** SOX9, **(B)** OLFM4, **(C)** CDX2, and **(D)** MUC5AC. White outline demarcates HP location. **(E)** MUC5AC staining for different GCHPs. **(F)** Low magnitude view of MVHPs and SSLs for MUC5AC staining. **(G)** MUC5AC staining of non-neoplastic colonic epithelium: (row 1) normal, (row 2) tumor adjacent, and (row 3) ulcerative colitis. **(H)** p-Creode trajectories inferred from epithelial scRNA-seq data. Overlay of genes relevant to developmental lineages. Node size represents the proportion of cells, and overlay color intensity represents mean scaled and standardized Arcsinh gene expression. **(I)** Additional RNA velocity maps for normal colonic tissues and polyps on combined UMAP embedding for (top) DIS and (bottom) VAL datasets. Vectors inferring average cell state transitions within each specimen are shown as black arrows. Colored points represent cells derived from the individual specimen, with grey points representing all other cells. **(J)** Representative genetic phylogenies inferred through DENDRO, with each tree representing a specimen for colonic polyp subtypes. Hierarchical clusters demarcated by dotted boxes and cell clusters denoted by leaf node color. Tree height is derived from a beta-binomial adjusted genetic divergence between single cells and shows the evolutionary time before splits in phylogeny. ABS – absorptive, ASC – adenoma specific cells, CT – crypt top, EE – enteroendocrine, GOB – goblet, STM – stem, SSC – serrated specific cells, TAC – transit amplifying cells, TUF – tuft.

To further leverage our single-cell transcriptomics data, we inferred cell-state transition trajectories from epithelial cells using p-Creode on batch-robust SCENIC regulons, which produced a developmental hierarchy where seven major cell types were mapped onto lineage branches [68] **(Figure 20E)**. From stem cells, we observed a secretory lineage branching into goblet, tuft, and enteroendocrine cells. Also, an absorptive lineage separately branched into proliferating transit-amplifying cells, colonocytes, and crypt top colonocytes **(Figure 21H)**. CytoTRACE score and WNT target gene overlays identified the stem cell branch **(Figure 20E; Figure 21H),** which was shared with ASCs, suggesting aberrantly expanded stem cells are the origin of AD. In marked contrast, SSCs, which expressed goblet cell genes such as *MUC2*, were inferred to develop from absorptive progenitors and colonocytes **(Figure 20E; Figure 21H)**. RNA Velocity analysis on individual tumors largely confirmed these findings **(Figure 20F; Figure 21I)** [83,84]. In normal colonic specimens, Velocity vectors originated from stem cells and flowed into differentiated cell types. ASCs were implicated to develop from stem cells following the RNA Velocity analysis. However, the Velocity vectors were reversed for SSCs, suggesting the origin of these cells to be non-stem cells.

Shared genetic variants between populations of neoplastic cells and normal cells can be used to deduce cellular origins. Because we sequenced normal cells within each polyp, we can leverage the continuum of mutational information between normal and neoplastic cells on a per polyp basis to determine their genetic distance and evaluate potential shared origins. To approximate the inherited genomic variations from single-cell data, we used DENDRO (DNA-based EvolutionNary tree preDiction by scRNA-seq technOlogy), a phylogenetic reconstruction algorithm that adjusts for the inherent sparsity of scRNA-seq data [81]. The robust application of this algorithm was enabled by the aggregation of transcriptionally similar cells and the filtering of variants using accepted quality metrics, which account for low sequencing depth on shorter reads and minimizing the inclusion of stochastically detected variants. Exonic variants detected through this

method were validated through exome sequencing of paired FFPE tissues. Representative DENDRO trees conducted on 34 libraries depicted the genetic variation across histological classifications and highlighted how ASCs had a shorter genetic distance to crypt base stem cells compared to SSCs (p<5E-02 MWU test) **(Figure 21J,K)**. In contrast, SSCs demonstrated divergent genetic profiles from stem cells, and, in fact, often clustered with differentiated colonocytes and absorptive progenitors **(Figure 21J)**. Orthogonal methodologies produced histological, transcriptomic, and genetic evidence to support the hypothesis that ADs arise from dysregulation of the stem cell compartment, but SSLs appear to arise from a developmentally committed cell.

Methods

Colorectal Molecular Atlas Project (COLON MAP) Cohort Recruitment and Characteristics

COLON MAP participants were recruited from among adults undergoing routine screening or surveillance colonoscopy or surgery for resection of a polyp at Vanderbilt University Medical Center in Nashville, TN, USA that began in March 2019 and is still on-going. The participants included in this study are the first 56 participants from COLON MAP with polyps collected for analysis by scRNA-seq. All participants provided written informed consent approved by the Vanderbilt University Medical Center Institutional Review Board.

Eligibility criteria for COLON MAP include ability to provide informed consent, free-living (not a resident of an institution), ability to speak and understand English, aged 40 to 75 years, permanent residence or telephone, and no personal confirmed or suspected histories of hereditary polyposis syndromes, familial or genetic colorectal cancer syndromes, inflammatory bowel disease, primary sclerosing cholangitis, colon resection or colectomy, cancer, neoadjuvant therapy, or cystic fibrosis. Eligible individuals were first identified from the schedule within the electronic health record (EHR) and assigned a random number. Potential participants undergoing colonoscopy

were further selected using a stratified weighted random sampling design to increase the inclusion of non-White or Latinx participants in the study. Within strata of colonoscopy appointment day and time, random sampling was weighted by EHR-derived racial/ethnic category (White non-Latinx vs all other races and ethnicities) such that non-White or Latinx patients were first selected at random within colonoscopy day and time. White non-Latinx patients were then selected at random within remaining time slots.

Following selection, study staff conducted a manual review of the EHR to confirm study eligibility. The majority of eligible individuals were mailed a letter to introduce the study and a few days later were attempted to be reached by telephone to discuss their willingness to participate in the study. Individuals who were willing to participate completed an additional screening form to confirm eligibility and eligible and willing individuals completed an interviewer-administered computer-assisted telephone interview to solicit information on personal health history, family history of cancer and polyps, lifestyle factors, and other risk factors for colorectal polyps and cancer. When the schedule of the study staff would allow, individuals who were not reached by telephone were approached in the colonoscopy waiting room or at the surgical appointment to determine eligibility and willingness as well as some individuals who did not receive a mailing.

For histopathological diagnosis, standard clinical histology was performed. Information on the colonoscopy or surgery and diagnosis was initially abstracted from the EHR colonoscopy, surgery, and pathology reports by study staff including *in vivo* size and polyp location. Two study pathologists additionally reviewed each case to standardize diagnoses and identify HP subtypes which are not part of routine clinical practice. For polyps which were partial due to the sampling for this study, the portion which had been reserved for clinical diagnosis was reviewed. SSLs were defined using the World Health Organization criteria of at least one distorted, dilated, or horizontally branched crypt within the polyp [173]. Subtypes of ADs were identified using standard

diagnostic criteria based on the villous component (tubular (< 25% villous component), tubulovillous (25%-74% villous component), and villous (≥ 75%)). HPs were classified as microvesicular HP or goblet cell HP [135]. In this analysis, participants were classified based upon the diagnosis of their index polyps but may have had synchronous polyps with the same or different histopathologies.

Cooperative Human Tumor Network (CHTN) Cohort Recruitment and Characteristics

Tissue was collected for COLON MAP from 33 colorectal cancer (CRC) patients via the CHTN Western Division. These participants were aged between 21 and 82 years of age from both sexes (51.5% male, 48.5% female) and were white (75.8%), Black (21.2%), or Asian (3.0%). De-identified clinical metadata from each patient was extracted from clinical pathology reports in accordance with policies from CHTN. Tumors were classified by grade and staging, ranging from G1 to G3 and I to IV, respectively. The majority (75.6%) of the tumors were classified as G2, or moderately differentiated, and staged primarily as IIA (30.3%) and IIIB (33.3%). Additionally, 51.5% were microsatellite stable (MSS) and 49.5% were microsatellite-high (MSI-H).

A colorectal carcinoma progression tissue microarray (TMA) was also provided by the CHTN Mid-Atlantic Division which included cores from 54 individuals. The mean (standard deviation) age of the individuals included on the TMA was 56.9 (14.7), 56.9% were men, and 43.1% were women. Race and ethnicity were not provided. Information on the TMA is available at https://chtn.sites.virginia.edu/chtn-crc2

Tennessee Colorectal Polyp Study (TCPS) Cohort Recruitment and Characteristics

The TCPS was a large colonoscopy-based case-control study among individuals undergoing colonoscopy in Nashville, Tennessee, USA between February 2003 and October 2010. Institutional approval for human subjects research was provided by the VUMC and VA Institutional

Review Boards and the VA Research and Development Committee. TCPS participants were aged between 40 to 75 years of age and had no personal history of colon resection, cancer, polyposis syndrome, inflammatory bowel disease, hereditary colorectal cancer syndromes, or previous adenoma. In TCPS, the diagnostic criteria for polyps were identical to the criteria used for COLON MAP. Additionally, all polyps were reviewed by one of the COLON MAP pathologists.

Detailed methods have been previously published [174]. In this analysis, a subset of TCPS formalin-fixed paraffin-embedded polyps which were previously analyzed by bulk RNA-seq were included to validate findings from the COLON MAP scRNA-seq analysis. In addition, a subset of fresh frozen polyps which were selected for targeted gene sequencing were also included.

COLON MAP Biological Specimen Collection and Processing, Colorectal Tissue

During the colonoscopy, the gastroenterologist used biopsy forceps to collect normal appearing mucosa samples from the ascending and descending colon for all participants. One of the biopsies from each colon segment was placed into RPMI. Any polyps were removed during the colonoscopy per standard clinical practice. In this analysis, the first polyp which was removed from a participant that was larger than 0.5 cm was selected for scRNA-seq analysis (index polyp). Polyps which were removed intact were bisected along the vertical access using a sterile razor blade and half was placed in RPMI. For polyps which were removed piecemeal, the second largest piece was placed in RPMI. The other portions of the polyps were placed into formalin for diagnosis and fixed and processed using standard clinical practice in the Vanderbilt Pathology Laboratory. All polyps which were placed in RPMI were immediately transported to the research lab for use in scRNA-seq analysis.

COLON MAP Bulk DNA Extraction

For germline, DNA was isolated from thawed buffy coat or mouth rinse samples using a QIAmp DNA kit (Qiagen). For tumors, DNA for whole exome sequencing (WES) was purified with the truXTRAC FFPE microTUBE DNA Kit-Column Purification kit (Covaris). In brief, tumor tissues were scraped from 1-5 of 10 µm FFPE sections, deparaffinized using xylene, and lysed in an optimized lysis buffer that contains proteinase K. Following the proteinase K digestion to release DNA from the tissue, a higher temperature was used incubation to reverse formalin crosslinking alongside RNase treatment using RNAse A (Thermo Fisher). The DNA and RNA samples were stored at -80°C before being used for assays.

CHTN Bulk DNA Extraction of Fresh Frozen Samples

Fresh frozen samples were stored in Tissue-Tek O.C.T. (Fisher Scientific) compound until ready for processing. These samples were washed in cold 1x PBS followed by centrifugation before using the Qiagen DNeasy Blood and Tissue kits (Qiagen) for DNA extraction. All following processing was performed according to the manufacturer's guidelines. The DNA extract collected from these samples were sequenced and aligned as detailed in the **COLON MAP Whole Exome Sequencing (WES) and Alignment** section.

TCPS Bulk DNA and RNA extraction

DNA was extracted from FFPE tissue sections using QIAamp DNA FFPE Tissue Kit (Qiagen), following the manufacturer's instructions. Briefly, tumor tissues were scraped from 1-5 of 10 µm FFPE sections, deparaffinized using xylene, and lysed under denaturing conditions with proteinase K. The sample lysate was incubated at 90°C to reverses formalin crosslinking and then applied to a QIAamp MinElute spin column, where DNA was captured on a silica membrane. The genomic DNA was then washed and eluted from the membrane.

DNA and total RNA were extracted from fresh frozen polyps and purified using Qiagen's AllPrep DNA/RNA/miRNA Universal Kit (Qiagen), following the manufacturer's instructions. Briefly, the frozen tissue samples were first disrupted and homogenized using Lysing Matrix E (MP Bio) by shaking the tubes on a bead-beater at 5.5 m/sec for 30 second. The lysate was then passed through an AllPrep DNA Mini spin column. This column allows selective and efficient binding of genomic DNA. Following on-column Proteinase K digestion, the column was then washed and pure, ready-to-use DNA was eluted. Flow-through from the DNA Mini spin column was then digested by Proteinase K in the presence of ethanol and applied to the RNeasy Mini spin column, where the total RNA binds to the membrane. Following DNase I digestion, contaminants were efficiently washed away and high-quality RNA was eluted in RNase-free water. The quantity and quality of the DNA/RNA samples were checked by Nanodrop (E260/E280 and E260/E230 ratio) and by separation on an Agilent BioAnalyzer.

COLON MAP Whole Exome Sequencing (WES) and Alignment

Standard WES was performed on S4 flow cells on NovaSeq6000 (PE150) to the targeted coverage. WES reads were aligned to the human reference genome hg19 using BWA [175], sorted and indexed by Sambamba [176]. Duplicated reads were removed by mark duplicates with Picard [177]. Somatic mutations were called using sequenced DNA extracted from specimens detailed in the **COLON MAP Biological Specimen Collection and Processing, Blood and Oral Rinse** section. These somatic mutations were then called using GATK4 Mutect2 in "normal-tumor" paired mode [178,179].

COLON MAP scRNA-seq, Single-cell Encapsulation and Library Generation

Colonic biopsy samples were first placed into RPMI solution, minced to approximately 4mm^2, and washed with 1x DPBS. These samples were then incubated in chelation buffer (4mM EDTA, 0.5 mM DTT) at 4 °C for 1 h 15 min. Then, the resulting tissue suspension was dissociated with cold protease and DNAse I for 25 minutes [180]. This suspension was titurated throughout the process,

every 10 minutes, then washed three times with 1x DPBS before encapsulation. Cells were encapsulated using a modified inDrop platform [19], and sequencing libraries were prepared using the TruDrop protocol [22]. Libraries were sequenced in a S4 flow cell using a PE150 kit on an Illumina NovaSeq 6000 to a target of 150 million reads.

COLON MAP scRNA-seq, Alignment and Droplet Matrix Generation

We demultiplexed, aligned, and corrected the detected read counts of these libraries with the DropEst pipeline [76], using the STAR aligner with the Ensembl reference genome [74], GRCh38 release 25. This was paired with the corresponding GTF annotations. The protocol for running this pipeline is described by [181].

COLON MAP scRNA-seq, Droplet Matrix Quality Control

We identified high-quality, cell-containing droplets and their respective barcodes through the joint application of cumulative sum inflection point thresholding, our dropkick QC algorithm [139], and prior-knowledge gene expression profiling. This droplet matrix was processed as an AnnData object using our preprocessing pipeline which utilizes the Scanpy toolkit [182]. First, we ran dropkick with 5-fold cross validation on the unprocessed droplet matrix, which assigned each barcode a probability of being a high-quality cell. Second, the droplet matrix was preprocessed for low dimensional analysis through finding the inflection point of the cumulative sum curve, and droplets with low information content were removed. Third, the remaining cells were normalized to the median number of counts per single-cell library per dataset, inverse hyperbolic sine transformed, and then scaled as a Z-score. Fourth, normalized matrices were projected into 2 dimensions by using its 50 principal component decomposition to initialize a UMAP [183]. Fifth, gene expression and dropkick probability scores were overlaid and checked for consistency. The genes overlaid were based on prior knowledge of the colonic epithelial markers, deferring to dropkick scores when no markers were found. Sixth, the selection of the final set of high-quality cell-containing

droplets were determined by setting a binarization threshold on the dropkick probability scores, given concordance to marker gene expression and other general quality metrics such as total counts, mitochondrial count percentage, and transcriptional diversity. The full protocol for running this QC pipeline is described by [181].

TCPS Targeted DNA Sequencing and Alignment

The list of candidate genes included in the targeted sequencing was developed from a literature review of candidate mutations which showed 1) evidence that mutation is common in adenoma (>5% of adenomas), 2) evidence that the mutation is associated with or predictive of adenoma recurrence in previous studies, 3) evidence that mutation is associated with clinically more significant adenoma (i.e. advanced adenoma or multiplicity), 4) evidence that mutation is associated with colorectal field carcinogenesis, and 5) evidence that mutation is associated with colorectal cancer aggressiveness and survival. In addition, additional candidate mutations were identified from potential mutations observed in Lrig1-Cre:Apc adenomas. All primer development and next-generation sequencing were conducted by Covance. Sequencing depth was 500X. Targeted sequencing reads were aligned to the human reference genome hg19 using BWA [175], and then were sorted and indexed by Sambamba [176]. Alignments were further refined, and variants were called using GATK Best Practices tools [179], including mark duplicates with Picard [177], base quality-score recalibration, and variant calling with HaplotypeCaller and GenotypeGVCFs [184]. SNPs were filtered using GATK VariantFiltration function with the parameters “QD < 2.0 || Qual < 30.0 || FS > 60.0 || SOR>3.0 || MQ < 40.0 || MQRankSum < -12.5 || ReadPosRankSum < -8.0”, while indels were filtered with the parameters “QD < 2.0 || Qual < 30.0 || FS > 200.0 || ReadPosRankSum < -20.0”. The variants with a minor allele frequency >0.1% in ExAC, gnomAD, TOPMed or 1,000 Genomes were also removed. The functional effects of variants were annotated by ANNOVAR [185,186].

TCPS Bulk RNA Sequencing and Alignment

Bulk RNA-sequencing was performed by Aros Applied Biotechnology A/S. This process involves the initial QC on an Agilent Bioanalyzer, with a minimum quality threshold of the DV_{200} at 30%. Total RNA-seq libraries which pass this QC threshold are prepared alongside a high-quality human reference RNA control. 100ng of RNA per sample is input to an Illumina TruSeq RNA Access Library Prep Kit, with protocol version 0.2. The yielded libraries undergo another round of QC through qPCR and quantified with a Qubit 2.0 Fluorometer, using its corresponding DNA BR Assay kit (Qubit), and size profiled on an Agilent Bioanalyzer. Pools of 4 libraries in equimolar amounts are created and undergo a final round of QC. These pools are loaded onto paired-end flow cells of a HiSeq2500 equipped with a cBot for sequencing at: 101 read cycles, 7 index cycles, and 101. The samples will be sequenced on a HiSeq2500 using 101 cycles for read 1, 7 index reads, and another 101 cycles for read 2. Following sequencing data generation, the reads are demultiplexed through Illumina's Genome Studio CASAVA software, which detected an average of 120 million reads per 4 sample pool.

scRNA-seq, Regulon Network Prediction, Activity Inference, and Visualization

The Single-Cell rEgulatory Network Inference and Clustering or SCENIC pipeline was used to integrate cancer, pre-cancer, and their corresponding normal tissue datasets [87,140]. For each group of integrated datasets, we concatenated the individual target datasets with an outer join and generated a combined AnnData object [182]. This AnnData object underwent further gene filtering, selecting only those that were expressed in at least 1% of all cells, primarily for the sake of speedup in running the module inference step of SCENIC. The resulting cumulative count matrix was input, without normalization, into the first step of SCENIC with default parameters, as suggested by the published protocol. We used a Dask client to parallelize the grnboost2 version of this step on an AMD Threadripper 2990WX CPU [187]. Subsequently, cisTarget was performed

using default parameters and three hg19 .feather ranking databases, comparing 10 species: tss-centered-5kb, tss-centered-10kb, and 500bp-upstream.

Further, this cisTarget step produced a list of detected regulons, their driving TFs, and their corresponding weights for the prediction of individual gene expression. These weights were used to build a feature matrix defining each regulon by its predicted targets. This feature matrix was then used to generate an adjacency matrix per SCENIC integration run, which was the basis of the regulon-regulon target network. This target network was based on a k-nearest neighbors graph (with k equal to the square root of the number of total regulons) of the adjacency matrix. For each of these target networks, the Louvain community detection algorithm was run at a resolution of 2, defining super-regulons [188]. This regulon-regulon target network (along with its cluster labels and average enrichment per regulon) was exported as a weighted adjacency matrix for visualization in Cytoscape [189].

Finally, we performed AUCell with default recommended parameters across 64 threads to generate a regulon activity enrichment matrix, which was jointly analyzed with the count-based matrix. Additional regulon activity enrichment scores were calculated for the Broad cohort by performing AUCell with regulon definitions learned from VUMC pre-cancer and CRC datasets. For visualization, target-network heatmaps featuring these regulon enrichment values were Z-score transformed, color scaled in a regulon-wise manner, and standardized to jointly integrated normal biopsies or polyp-derived normal cells when possible.

scRNA-seq, Count Matrix Normalization and Heatmap Generation

Using scanpy and numpy functions, raw count data were normalized by median library size, log-like transformed with Arcsinh, and Z-score standardized per gene [182,190]. This yielded interpretable

unit variance scaled and centered values. Heatmaps featuring individual gene expression depict this normalized, transformed, and standardized data with color scaling in a gene-wise manner.

scRNA-seq, UMAP and t-SNE Visualization

Three modes of UMAP visualization were used in this study based on regulons, feature-selected counts, or Harmony-corrected components. All human epithelial UMAP visualizations were generated using the "scanpy.tl.umap" function with a min_dist parameter of 0.15. The input to this function was Z-score standardized AUCell values, their 50-principal component decompositions with no feature selection, and a subsequent KNN graph with k equal to the square root of the number cells projected. Human nonepithelial UMAP visualizations that included all nonepithelial subtypes were performed the same way. To finely resolve T cell subtypes with UMAP, we generated a KNN based on the PCA of a feature-selected set of genes after normalizing, log-like transforming with Arcsinh, and Z-score standardizing raw counts. Finally, murine validation experiments were integrated with the Harmony algorithm, generating adjusted principal components with default parameters [191]. These components were used as the basis for KNN and UMAP generation with the same parameters as used in the human data. For t-SNE visualizations, the perplexity was set to the same as the k used in the UMAP KNN graph. The bootstrapped variant of t-SNE visualization was performed by running t-SNE with the same parameters 100 times to ensure qualitatively robust embeddings, given the algorithms inherent stochasticity.

scRNA-seq, Gene Signature Scoring

We used a gene signature scoring method implemented in scanpy and first detailed by (Satija et al., 2015). This method scores a defined gene set by finding the difference between its average expression against the average expression of randomly sampled sets of reference genes, corresponding to matched and binned expression levels. Each signature in this study was

calculated on normalized, transformed, and standardized data (as described in the **scRNA-seq, Count Matrix Normalization and Heatmap Generation** section) using a reference sample size of 2000 genes across 25 bins. The x-axis range of scatterplots featuring these signature data was set by excluding single-cell outliers beyond the 1.5x interquartile range [192,193]. Statistical tests of these score distributions encompass an initial Kruskal-Wallis test. If the null hypothesis was rejected, these tests were followed by post-hoc Mann-Whitney U tests and appropriate p-value adjustments [194,195]. Gene signatures for murine TSC scRNA-seq were calculated for ISCI, ISCII, and ISCIII as described by Biton et al, with the same method applied to calculating the murine MHCII signature [196].

scRNA-seq, Unsupervised Clustering and Cell Type Labeling

The labeling of single-cell subpopulations was done through the Leiden algorithm, as part of the Scanpy toolkit. We performed Leiden clustering based on the KNN derived from the distances calculated in the principal component space of Z-score transformed regulon enrichment scores, as these represented cell-cell transcriptional states in a more batch-robust manner. The resolution of this clustering was based on the detection of rarer populations such as enteroendocrine cells, at 2. Since this algorithm detected discrete clusters in a continuum of cell states, we aggregated multiple discrete clusters by the observation of marker gene expression. Similarly, these methods were applied to nonepithelial datasets given their regulon or feature-selected matrices, depending on the subtypes of interest. This Leiden algorithm was also used to determine clusters for murine scRNA-seq validation experiments. Higher resolution subclustering was also done by performing k-means clustering after the initial Leiden clustering. Importantly, some subclusters were identified as a result of patient-to-patient variation originating from mitochondrial read enrichment, as evidenced by mitochondrial read percentage distributions and GO terms. These subclusters were identified and statements regarding their relative, subpopulational variation were excluded. These patient-to-patient variations did not affect overall comparisons between tumor-specific and normal

cell types. For example, after excluding these mitochondrially-enriched subclusters, the SSC subpopulational analysis focused on GO terms related to intercellular communication and stromal interactions.

scRNA-seq, Differential Gene Expression Testing and Gene Set Enrichment Analysis (GSEA)

The differential testing of gene expression was performed based on cluster labels (as defined by the **scRNA-seq, Unsupervised Clustering and Cell Type Labeling** section), both in the context of raw gene counts and regulon enrichment values. For both cases, we used Mann-Whitney U tests with Benjamini-Hochberg corrections, on the raw values, implemented through the "scanpy.tl.rank_genes_groups" function, identifying the top 200 genes and top 50-100 regulons [182]. Further, biological insight was gathered through scanpy's integration of g:profiler gene set enrichment framework [165]. This process was also performed on the stem and TSC components of the murine scRNA-seq datasets using the GSEA webapp [197,198].

scRNA-seq, Proportional Cell Type Representation and Identifying Polyp-Specific Populations

Given the detected clusters (as described in the **scRNA-seq, Unsupervised Clustering and Cell Type Labeling** section), we calculated the proportional cell type representations of each individual sample. We counted the raw number of epithelial and nonepithelial cells as well as the raw number of cells falling into any given cell cluster. These results were cross-tabulated as contingency tables, summarizing how many cells were observed in each category and for which samples using pandas [199]. Proportional values were then calculated by normalizing cluster counts to the number of epithelial cells per sample **(Figures 2 and S2)** or to the cumulative number of cells per sample **(Figures 6 and S6).** Clusters were designated as polyp-specific populations if, proportionally, they were significantly overrepresented in polyp samples and not normal samples,

which was indicated by post-hoc statistical tests following Kruskal-Wallis null hypothesis rejection. The x-axis range of scatterplots featuring these proportional representation data was set by excluding samples with values beyond the 1.5x interquartile range. In the context of the murine scRNA-seq datasets, the proportional representation of cell types was calculated by normalizing to the total number of epithelial or immune cell subtypes for each Mist1 and Lrig1 tumor sample.

scRNA-seq, Predicting Differentiation Potential with CytoTRACE

CytoTRACE is a relative scoring method dependent on included datasets for inferring developmental potential. CytoTRACE was performed based on the default recommended settings after concatenating the batches of interest using an outer join [144]. We performed CytoTRACE with five separate groupings of single-cell libraries. First, the discovery cohort **(Figures 2E, S2E),** including all epithelial cells from both its normal biopsies and polyps. Second, the validation cohort **(Figures 2E, S2E)**, including all epithelial cells from both its normal biopsies and polyps. Third, the epithelial VUMC polyp-specific cells **(Figure 4E)**, including only tumor-specific cells from VUMC AD, MSS, SER, and MSI-H samples. Fourth, the epithelial Broad cohort **(Figures 4E,S4E)**, including MSS, MSI-H, and Normal samples. The Broad cohort (including 32 normal samples) distribution was only calculated from 50% random sample of the total cells detected due to memory constraints. Fifth, CytoTRACE was performed on the stem and TSC component of the murine scRNA-seq datasets. Statistical tests of these score distributions encompass an initial Kruskal-Wallis test. If the null hypothesis was rejected, these tests were followed by post-hoc Mann-Whitney U tests and appropriate p-value adjustments.

scRNA-seq, CMS scoring at Single-cell Resolution

The single-cell distributions of CMS scores were calculated on the VUMC ASC, MSS, SSC, and MSI-H and the Broad MSI and MSI-H libraries using the CMSclassifier R package as described by [128,200]. To accommodate the heterogeneity of the single-cell landscape, the single sample

predictor or SSP mode of the software was used after converting gene symbols to Entrez IDs. This SSP mode calculated the median correlation distance between each single cell to established, standard centroids derived from CMS1, CMS2, CMS3, and CMS4 CRC subtypes. Further, these score distributions were visualized through a normalized kernel density estimation implemented in the Seaborn python package. Statistical tests of these score distributions encompass an initial Kruskal-Wallis test. If the null hypothesis was rejected, these tests were followed by post-hoc Mann-Whitney U tests and appropriate p-value adjustments.

scRNA-seq, Trajectory Inference

pCreode was used to map the developmental state transitions of the single-cell transcriptional landscape of our Discovery cohort pre-cancer and normal COLON MAP samples [68]. This algorithm was generalized to process regulon-based principal components, inheriting its batch-robust properties. By examining the variation captured by the principal components, we selected the first 4 components based on their capture of rare cell populations, such as Tuft and enteroendocrine cells. We developed this algorithm to traverse a density weighted KNN generated from the pairwise distances between each single cell; subsequently, we used a histogram thresholding method to estimate the neighborhood distance cutoff for calculating local densities. These densities were used as input to a supervised variant of pCreode, which established developmental endstates through K-means clustering and marker-defined labels. The downsampling and noise parameters were both set to 4, resulting in samples of around 6,000 cells per run, and repeated 50 times. Each of these runs was scored by the minimization of the Gromov–Hausdorff distance, resulting in a single, most representative graph layout. Overlays were generated based on pre-computed single-cell observation vectors, such as a CytoTRACE score, or the normalized, transformed, and z-scored gene expression values.

scRNA-seq, RNA Velocity

RNA velocity analysis was performed using velocyto CLI version 0.17 [201]. Individual sample BAM files were used as input to the "run-dropest" command along with a human gene annotation file (GTF) for GRCh38.85, and a tab-delimited text file containing dropkick-filtered cell barcodes from the corresponding sample as the "—bcfile" flag. Then, scVelo version 0.2.3 was used to build models of splicing kinetics to estimate and visualize RNA velocity vector fields in SCENIC integrated UMAP space [84]. Each sample was individually filtered to the top 2,000 genes expressed in a minimum of 20 cells using the function "scvelo.pp.filter_and_normalize". The moments of all RNA velocity vectors were calculated with 30 principal components and 30 nearest neighbors using the function "scvelo.pp.moments" prior to estimating velocities using "scvelo.tl.velocity" with default parameters. Finally, velocity UMAP embeddings were plotted using the function "scvelo.pl.velocity_embedding_stream" and the subset of SCENIC master UMAP coordinates for each sample.

sc-RNA-seq, Subclone Phylogeny Estimation

We used DENDRO, an algorithm designed to reconstruct subclonal phylogenies within scRNA-seq datasets [81]. Since our sequencing libraries are generated through the tag-based inDrop method, the short, 3'-biased reads necessitated the aggregation of single-cell transcriptomes. For each of the 34 sequencing libraries we performed this analysis on (24 ADs, 11 SERs), we defined 20 aggregate populations through regulon-based K-means clustering [202]. Thus, we predicted the average genotypic representation of multiple single-cells, or pseudo-bulk RNA-seq libraries, and created a phylogenetic tree between the defined clusters. DropEst produced a filtered and sorted .bam file, which we derived read information from, and split into 20 distinct .bam files using the sinto python package [203]. These bam files were then processed with GATK4 [179], according to guidelines detailed by Zhou et al. The GATK4 steps of this pipeline involved the following: adding read groups, marking duplicates, splitting N-cigar reads, applying base quality recalibration with known single nucleotdie polymorphisms (SNPs), haplotype calling (with the GATK4

HaplotypeCaller and an hg38 reference) to generate VCF files, consolidating these VCFs into genomicsdb databases, and then genotyping these data.

Because the measurement of single nucleotide variations (SNVs) within transcriptomes is dependent on dynamic expression patterns, we used a beta-binomial framework, as described by Zhou et al., to model genetic divergence between each pseudo-bulk, cell aggregate. Standard genetic divergence frameworks, such as those comparing DNA-derived genomic variants, do not consider the varying levels of low nor high gene expression between pseudo-bulk populations. Examples of this transcription-specific variation would be stochastic bursts of gene expression captured in a minority of populations, yielding low average expression across all populations, and constitutively expressed genes, yielding high average expression across all populations. These bursty loci will more likely represent genes dropped out from the majority of pseudo-bulk populations, so including its respective variants would yield an inflated genetic divergence value. Conversely, variants in loci that are expressed and observed in the vast majority of pseudo-bulk populations would be uninformative in terms of phylogenetic discrimination. The genetic divergence *d* between each possible cell aggregate pair c and c' at loci *g* is represented formally as:

$$d_{cc'}^{g} = \log \frac{P\left(X_{cg}, X_{c'g} \mathrm{I} N_{cg}, N_{c'g}\right)}{P\left(X_{cg}, X_{c'g} \mathrm{I} N_{cg}, N_{c'g}, I_c = I_{c'}\right)}$$

Where *c* and *c'* represent two different cell aggregates, while *I* and *I'* represent their originating clonal groups. Correspondingly, X_{cg} is the alternative allele read count for cell aggregate *c* at loci *g*, while N_{cg} is the respective total read count. Thus, *d* is a function of five derived probabilities:

$$P_g$$

1. Which is the alternative allele frequency across cell aggregates estimated by the above GATK4 pipeline.

$$P(X_{cg}\mathrm{IN}_{\mathrm{cg}}, Z_{cg} = 0)\ and\ P(X_{c'g}\mathrm{IN}_{c'g}, Z_{c'g} = 0)$$

2 and 3. Which represent detected variants due to rare editing and technical sequencing error events in *c* and *c'* at *g*. Here, Z_{cg} is set as 0, modeling scenarios lacking SNVs, which can be approximated as the following binomial distribution with ϵ set to 0.001 or 0.1%.

$$P(X_{cg}\mathrm{IN}_{cg}, Z_{cg} = 0) \sim\ Binomial(X_{cg}\mathrm{IN}_{cg}, \varepsilon)$$

ϵ, representing the combined error rate, was used according to our sequencing platform, a NovaSeq 6000 System. This is in line with empirical studies of Illumina sequencing instruments as detailed by Stoler et al., observing a median error rate of 0.109% across 239 samples on a NovaSeq 6000 device [204]. Another previous study by Fox et al. had similar estimates for sequencing-by-synthesis platforms, including the Illumina MiSeq and HiSeq2000, with an error frequency of 10^{-3} (or 0.1%) attributed to single nucleotide substitutions [205].

$$P(X_{cg}\mathrm{IN}_{\mathrm{cg}}, Z_{cg} = 1)\ and\ P(X_{c'g}\mathrm{IN}_{c'g}, Z_{c'g} = 1)$$

4 and 5. Which represent detected variants due to the presence of SNVs in *c* and *c'* at *g*. In this case, Z_{cg} is set as 1, modeling scenarios with SNVs present.

$$P(X_{cg}\mathrm{IN}_{cg}, Z_{cg} = 1) \sim\ \int_0^1 Binomial(X_{cg}\mathrm{IN}_{cg}, Q_{cg} = q)\ dF(q),\ q\ \sim\ Beta(\alpha_g, \beta_g)$$

This can be approximated as a beta-binomial distribution, as previously described by Jiang et al. and Skelly et al. in the context of single-cell and bulk RNA sequencing [206,207]. Q_{cg} is the proportion of alternative alleles in cell aggregate *c* at *g*, using a beta distribution prior, approximated as *q*. *q* is parameterized by α_g and β_g, as estimated gene activation and deactivation rates respectively.

Before performing genetic divergence calculations based on these probabilistic models, two filters were applied to minimize the inclusion of stochastically or constitutively expressed variants:

The first filter is dependent on the observed variant allele frequencies (VAFs) across each set of cell aggregates. VAFs were visualized as histograms representing the number of times each unique variant was observed across each set of cell aggregates. We observed that these VAF distributions were unimodal and positively skewed, with the vast majority of variants being detected in very few cell aggregates, which was in line with stochastic gene expression. To remove these stochastically expressed variants, we heuristically determined a cutoff at observed convex elbow/knee points of the curve, at 10%. This cutoff was symmetrically applied to the top 10% of the most pervasive variants as well, as these represented constitutively expressed variants.

The second filter is dependent on α and β parameter estimations. If either the α or β parameters of the beta prior were estimated to be 0 or 1, it meant that the activation and deactivation rates were completely on or off. Akin to the rationale for our first filter, these variants would not be informative in the genetic divergence calculation since they likely represent genes with a tendency to dropout/be stochastically expressed or be constitutively active. These cases would inflate or deflate genetic divergence metrics, respectively.

The quality of the filtered variants, consisting of about 5.07% (std. 1.46%) of the initially detected variants, met appropriate QD and DP levels suggested by GATK4 guidelines and were also located within genomic regions characteristic of the inDrop barcoding chemistry (https://github.com/Ken-Lau-Lab/STAR_Methods/blob/main/Supplemental_Table_Variant_Type_Func.refGene_Distribution.xlsx). Exonic variants detected through this method were validated through the exome sequencing

of paired FFPE tissue and respective GATK HaplotypeCaller pipeline. If the exact exonic genomic loci and genotypes were detected in both the exome and scRNA-seq pseudo-bulk aggregates, the variants were flagged as validated. An average of 53.9% (std. 12.9%, max >75%) of the exonic variants detected through scRNA-seq were validated with this orthogonal exome sequencing. Tables of these detected variants per cell population and their exome-seq statuses are shown at (https://github.com/Ken-Lau-Lab/STAR_Methods/tree/main/Tables).

After filtering, the genetic divergence is calculated for all possible pairs of cell aggregates, and a phylogenetic tree is constructed. The leaves of these trees represent the previously defined cell aggregates, which were assigned cell type labels accordingly. For each set of pseudo-bulk cell aggregates, we also calculated the minimum genetic divergence between tumor-specific cell aggregates (ASCs and SSCs) and canonical stem cell aggregates (STM). These values were normalized to the maximum distances observed per tumor sample, yielding a value between 0 and 1. This metric was interpolated with a value of 1 in samples which lacked measurable canonical stem cell aggregates.

Chapter IV - Multi-modal data integration and model validation through organoids and genetically engineered mouse models

Recreated from:

Chen, B., Scurrah, C. R., Mckinley, E. T., Simmons, A. J., Ramirez-Solano, M. A., Zhu, X., Markham, N. O., Heiser, C. N., Vega, P. N., ... Coffey, R., Shrubsole, M., & Lau, K. S. (2021). Differential pre-malignant programs and microenvironment chart distinct paths to malignancy in human colorectal polyps. Cell. https://doi.org/10.1016/j.cell.2021.11.031

Introduction

Without experimental validation, *in silico* computational models remain entirely theoretical and observational. Clearly, the vast majority of computationally generated models are not directly testable, since, outside of clinical trials, the *in vivo* validation of models of human biology remains infeasible. Thus, the utilization of human organoids *ex vivo* and mouse models *in vivo* are the most feasible means of experimental validation for generalized models described in humans. Here, we present in-depth dissections of the malignant transformation of human precancer followed by the experimental support of the hypothesized effects of divergent tumorigenesis originating from variations in cell differentiation state.

Phenotypic transitions and subtype-specific features during malignant progression from pre-cancer to cancer

We used the same experimental design of our polyp studies to examine MSI-H and MSS CRCs in relation to their precursors. We performed scRNA-seq on 7 (2 MSI-H, 5 MSS) fresh CRC specimens for a discovery dataset and procured a CRC scRNA-seq dataset (n=60; 32 MSI-H, 28 MSS) from the Broad Institute. Furthermore, we analyzed whole tumor blocks from 26 additional CRC patients (14 MSI-H, 12 MSS) for further validation. Exome sequencing of CRC specimens

revealed expected mutational features in MSS CRCs following the conventional tumorigenesis pathway with *APC* (100%), *KRAS* (35%), and *TP53* (71%) mutations **(Figure 23A)**. MSI-H CRCs had fewer of these conventional mutations (33%,0%,7%, respectively), but more *BRAF* mutations (53% in MSI-H vs. 0% in MSS). All MSI-H CRCs had acquired a higher mutational burden compared to MSS CRCs. Histologically, all CRCs showed invasive adenocarcinoma with cribriform architecture **(Figure 23B)**, with MSI-H CRCs exhibiting mucinous features.

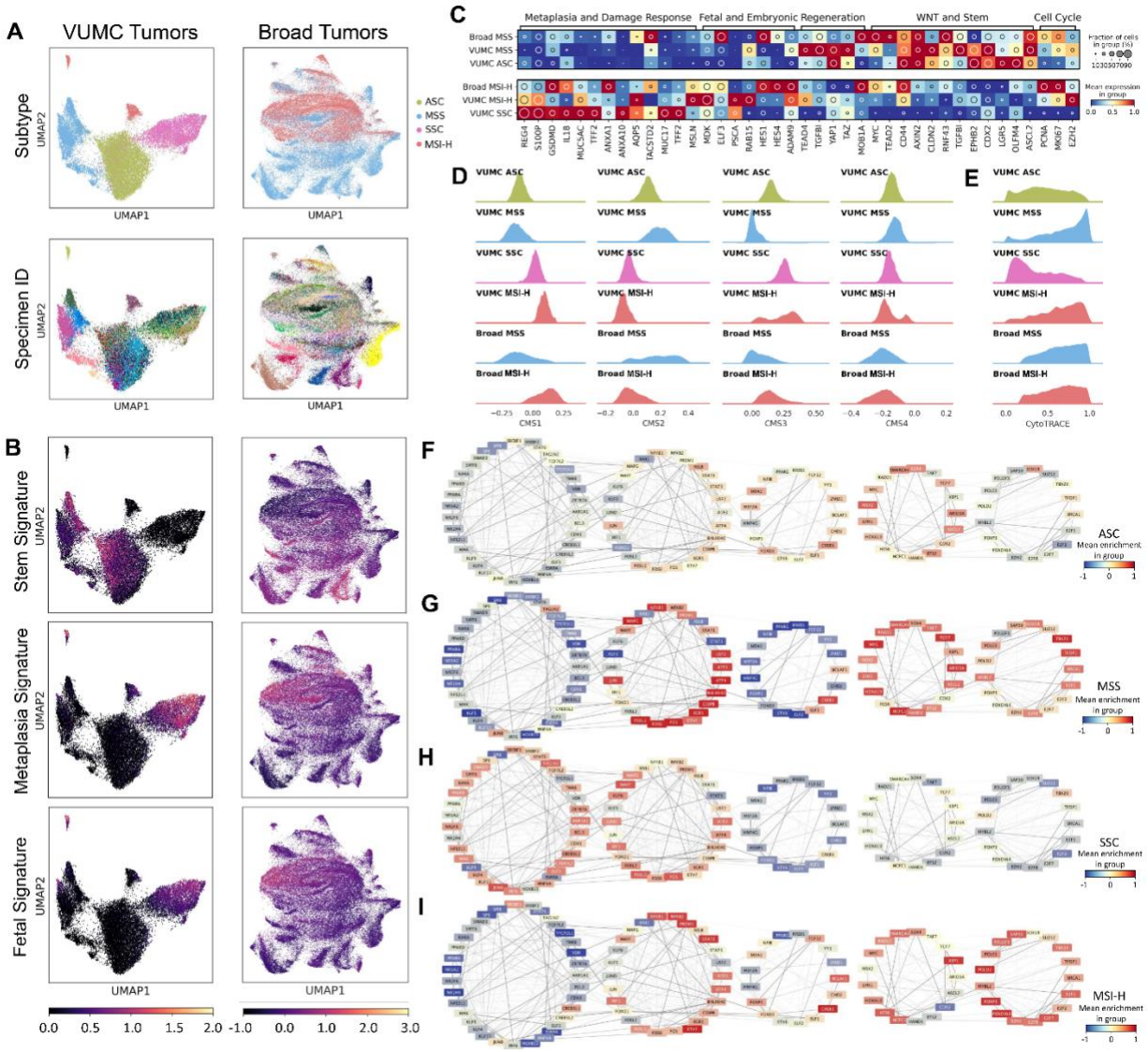


Figure 22. Analysis of CRCs through the lens of pre-cancers.

(A) Regulon-based UMAPs for tumor-specific cells derived from pre-cancerous and cancerous specimens colored by (top) subtypes and (bottom) specimen ID for both the (left) VUMC and (right) Broad datasets. **(B)** Stem, meta, and fetal signature scores overlaid onto UMAP of tumor-specific cells for both (left) VUMC and (right) Broad datasets. **(C)** Heatmap representation of pre-cancer derived gene sets for VUMC (n = 55) and Broad (n = 60) tumor-specific cells. Inset circle indicates prevalence within defined single-cell populations and color intensity represents scaled and standardized Arcsinh gene expression. **(D)** Single-cell CMS scoring based on single sample predictor for both VUMC and Broad tumor-specific cells. **(E)** Ridge plots of CytoTRACE score distributions for VUMC and Broad tumor-specific cells. Broad CytoTRACE scores calculated relative to corresponding Broad normals (n=32) **(F-I)** TF target network created from polyp-specific and cancer cells, organized into super-regulons derived from clustering of shared targets: **(F)** ASC, **(G)**, MSS, **(H)** SSC, and **(I)** MSI-H. Color overlays for each TF node are averaged and normalized regulon enrichment scores, while edge opacities are the inferred TF-target weightings.

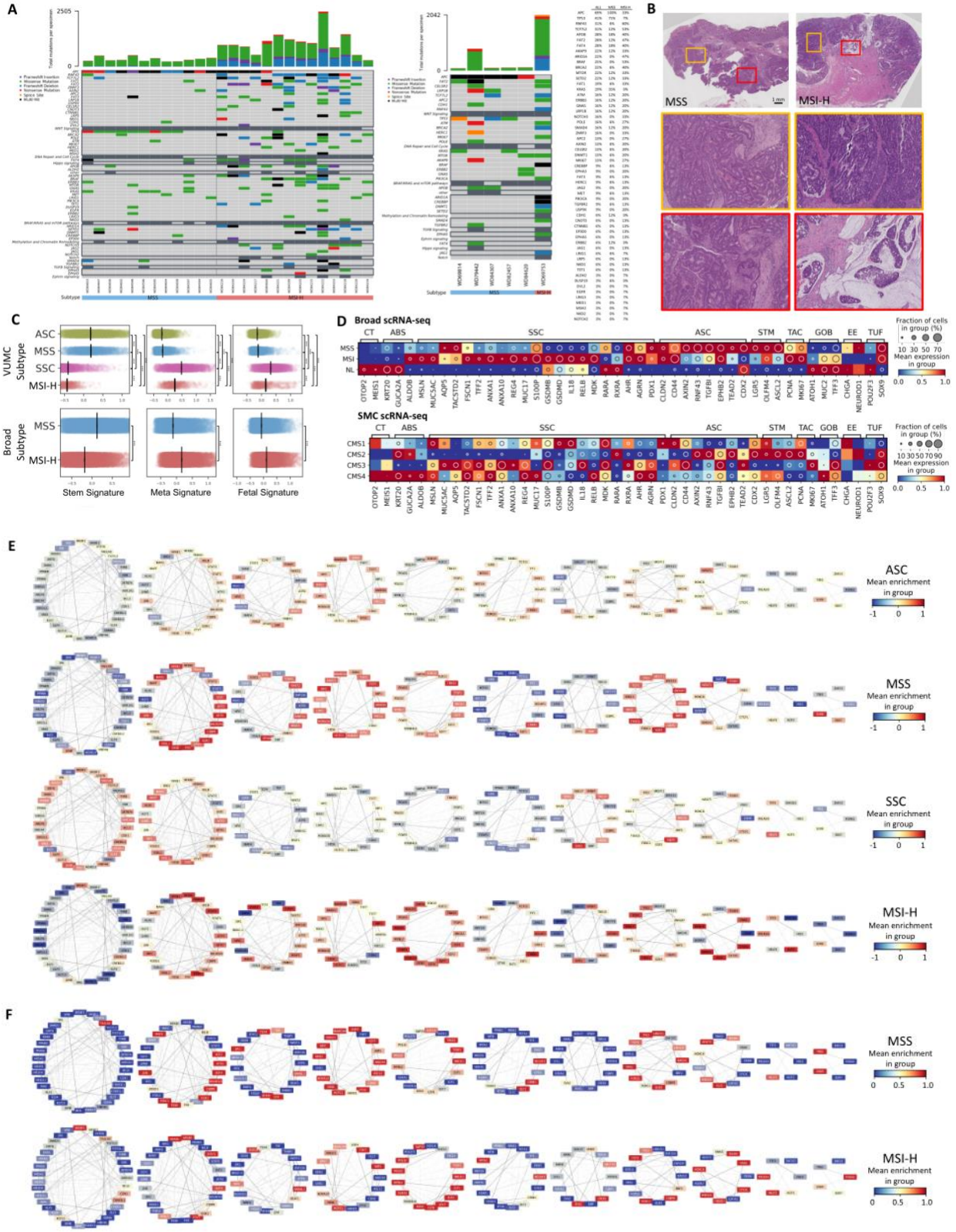

A
B
MSS
MSI-H
C
VUMC Subtype
ASC
MSS
SSC
MSI-H
Broad Subtype
MSS
MSI-H
Stem Signature
Meta Signature
Fetal Signature
D
Broad scRNA-seq
SMC scRNA-seq
CT
ABS
SSC
ASC
STM
TAC
GOB
EE
TUF
Fraction of cells in group (%)
Mean expression in group
E
ASC
MSS
SSC
MSI-H
Mean enrichment in group
F
MSS
MSI-H
Mean enrichment in group

Figure 23. Single-cell characterization of CRC subtypes as related to their pre-cancer precursors.

(A) Oncoplot representation of the mutational landscape of CRCs detected through exome sequencing, and (left) mutational variant calling or (right) somatic mutation calling. Total number of mutations detected per specimen represented as bar plot (top), and different type of mutations color-coded. Important genes to CRC are presented and grouped into pathways. Proportion of gene and pathway mutation within CRC subtypes (combined) summarized as table. **(B)** Representative H&E of each CRC subtype at low and high magnification views of the corresponding colored insets. **(C)** Scatter plots of signature scores by tumor-specific cell subtype, with each point representing a single cell, for (top) VUMC and (bottom) Broad datasets. Error bars depict SEM of single cells. **(D)** Heatmap representation of top biologically relevant and differentially expressed genes, for Broad (MSS, n=28; MSI-H, n=32) and SMC (CMS1, n=5; CMS2, n=8; CMS3, n=4; CMS4, n=6) single-cell datasets. Inset circle indicates prevalence within defined single-cell populations and color intensity represents scaled and standardized Arcsinh gene expression. **(E,F)** Expanded TF target network created from **(E)** VUMC or **(F)** Broad tumor-specific cells, organized into super-regulons derived from clustering of shared targets. Color overlays for each TF node are averaged and normalized regulon enrichment scores, while edge opacities are the inferred TF-target weightings. VUMC regulon values are normalized to tumor-derived normal cells while Broad values are unnormalized due to lack of detected tumor-resident normal cells.

scRNA-seq data of malignant CRC cells revealed substantial tumor-to-tumor variability, as seen by others [93,208]. Transcriptomes clustered by individual tumors even after regulon-based embedding **(Figure 22A)**. Combined with our pre-cancer data, an increase in intertumoral heterogeneity was observed as epithelial cells transition from normal to pre-cancer to malignant cells. We considered that the intrinsic complexity and heterogeneity of CRC transcriptomics might be reduced by looking at CRC cells through the lens of pre-cancerous polyps; by using the pre-identified gene sets from ADs and SERs, we observed that both MSS and MSI-H CRC cells retained aspects of their respective precursors. Comparing between subtypes, MSS CRC cells overexpressed a signature of regenerative crypt base stem cells, and MSI-H CRC cells retained a metaplastic signature **(Figure 22B,C; Figure 23C).** These patterns of gene expression in different CRC subtypes were replicated using scRNA-seq data generated independently from the Samsung Medical Center (SMC) [208] **(Figure 23D).** To further support commonalities between pre-cancer and cancer, we classified ASCs, SSCs, and CRC cells by consensus molecular subtype (CMS), which uses the median distance between transcriptional centroids of each CMS to individual transcriptomes as a single-cell predictor [128,200] **(Figure 22D)**. ASCs shared comparable score distributions with MSS CRC cells, which were predicted to be CMS2 **(Figure 22D)** the subtype most often associated with *APC* mutations and WNT pathway dysregulation. In contrast, both SSCs and MSI-H CRC cells scored low for CMS2, but high for CMS1 and CMS3, which feature immunogenic and RAS pathway activation, respectively [209–211] **(Figure 22D)** None of the examined polyp cells had strong enrichment of CMS4 epithelial-to-mesenchymal transition scores **(Figure 22D),** consistent with their identity as early tumor cells and the previously reported absence of CMS4 in ADs [212,213]. Shared features between malignant cells and pre-cancerous cells provide additional evidence of precursor-cancer relationships.

In addition to commonalities in CRCs and their precursors, we also examined characteristics acquired or lost during the transition from pre-cancer to malignancy. MSI-H CRC cells showed

relatively decreased metaplastic and fetal features compared to SSCs. However, key genes within the WNT-activated stem cell program were increased relative to SSCs **(Figure 22C; Figure 23C).** Supporting reactivation of stem cell properties, CytoTRACE analysis demonstrated MSI-H CRC cells had higher inferred stem potential compared to SSCs **(Figure 22E; Supplemental Table S5)**. MSS CRC cells also gained higher scores relative to ASCs, suggesting they possess aberrant stemness (**Figure 22E)**. Gene regulatory network analysis more clearly demonstrated how molecular pathways that were either maintained or were altered during malignant transition, supported through GSEA **(Figure 22F-I; Figure 23E)**. A common feature for both CRC subtypes was the activation of the proliferation super-regulon, with enrichment of genes (*PCNA* and *MKI67*) and regulons (BRCA1, RAD21, POLE3, and EZH2) involved in DNA synthesis and repair **(Figure 22F-I)**. The WNT signaling super-regulon was consistently upregulated in ASCs and MSS CRC cells **(Figure 22F,G)**. For MSI-H CRC cells, the super-regulon describing pathogen damage response in SSCs was suppressed**,** but the WNT signaling super-regulon, previously suppressed in SERs, was activated **(Figure 22H, I)**. The differences in super-regulon enrichment were maintained in the Broad dataset **(Figure 23F)**. Activation of the WNT pathway was supported by acquisition of activating mutations in non-APC WNT pathway components in MSI-H CRCs, including *RNF43* (60%), *TCF7L2* (53%), ZNRF3 (33%), *APC2 (27%)*, *AXIN2* (20%), *FAT1* (33%), *FAT2* (47%), and *FAT4* (40%) **(Figure 23A)**. TCGA exome sequencing data also showed enrichment of non-APC WNT pathway gene mutations in MSI-H CRC (*APC2*, *RNF43*, *AXIN2*, *LRP1B*, *LRP6*, *TCF7L2*). **(Figure 25A)** [214]. These results suggest MSI-H CRC acquired metaplasia-independent events by transitioning into more aggressive stem-like cells through selection of APC-independent activating mutations in the WNT pathway.

Transition of metaplastic cells to stem-like cells contributes to tumor heterogeneity in MSI-H CRCs

We further queried 63 annotated bulk RNA-seq datasets from the TCGA and observed consistent expression of a gastric metaplastic signature in CMS1 and CMS3 CRCs but not CMS2 CRCs **(Figure 25B)**, and the inverse was true for the stem cell signature. However, the data were noisier than scRNA-seq data on an individual tumor basis, likely due to additional intratumoral heterogeneity and/or poor data quality (Muzny et al., 2012). This led us to perform spatial profiling using multiplex immunostaining and whole slide scanning of entire CRC specimens. Strikingly, none of the MSS CRCs (0/17) stained positive for MUC5AC, but most MSI-H CRCs (13/14) had some degree of staining **(Figure 24A,B)**. However, the amount of tumor area stained by MUC5AC was variable within the positive MSI-H CRCs. CDX2 staining followed the inverse trend; virtually all tumor cells in MSS CRCs were CDX2-positive, and MSI-H CRCs had variably decreased CDX2 staining **(Figure 24A,B)**. Stem cell markers (OLFM4, SOX9) were expressed throughout MSS CRCs, and they uniformly lacked MUC5AC expression **(Figure 24C,D; Figure 25C)**. In contrast, MSI-H CRCs displayed considerable intratumoral heterogeneity. Unlike SSLs where almost all crypts were MUC5AC-high, MUC5AC staining was variable and low in certain regions of MSI-H tumors **(Figure 24E,F).** Interestingly, these MUC5AC-low regions were positive for OLFM4 and to some degree CDX2 **(Figure 24F-H)**. SOX9 was overexpressed in MSI-H CRCs in both MUC5AC-high and -low regions, suggesting all cells gained some level of stem-like characteristics upon malignant progression **(Figure 24H)**. We validated the heterogeneity between stem and metaplastic cells by focused analysis of individual scRNA-seq datasets from MSI-H CRCs. Positive *MUC5AC* and *MSLN* expression, coupled to loss of *CDX2* expression, distinguished metaplastic cells from *LGR5*/β-catenin-expressing, WNT-driven stem cells within the same tumor **(Figure 24I)**. These stem-like cells were enriched in cell cycle gene expression consistent with the increased proliferative capacity of CRC cells identified in our earlier analysis **(Figure 24I, 4C,E)**. In multiple instances of MSI-H CRCs, we observed intratumoral cellular heterogeneity characterized by the mutual exclusivity of stem-like cells and metaplastic cells.

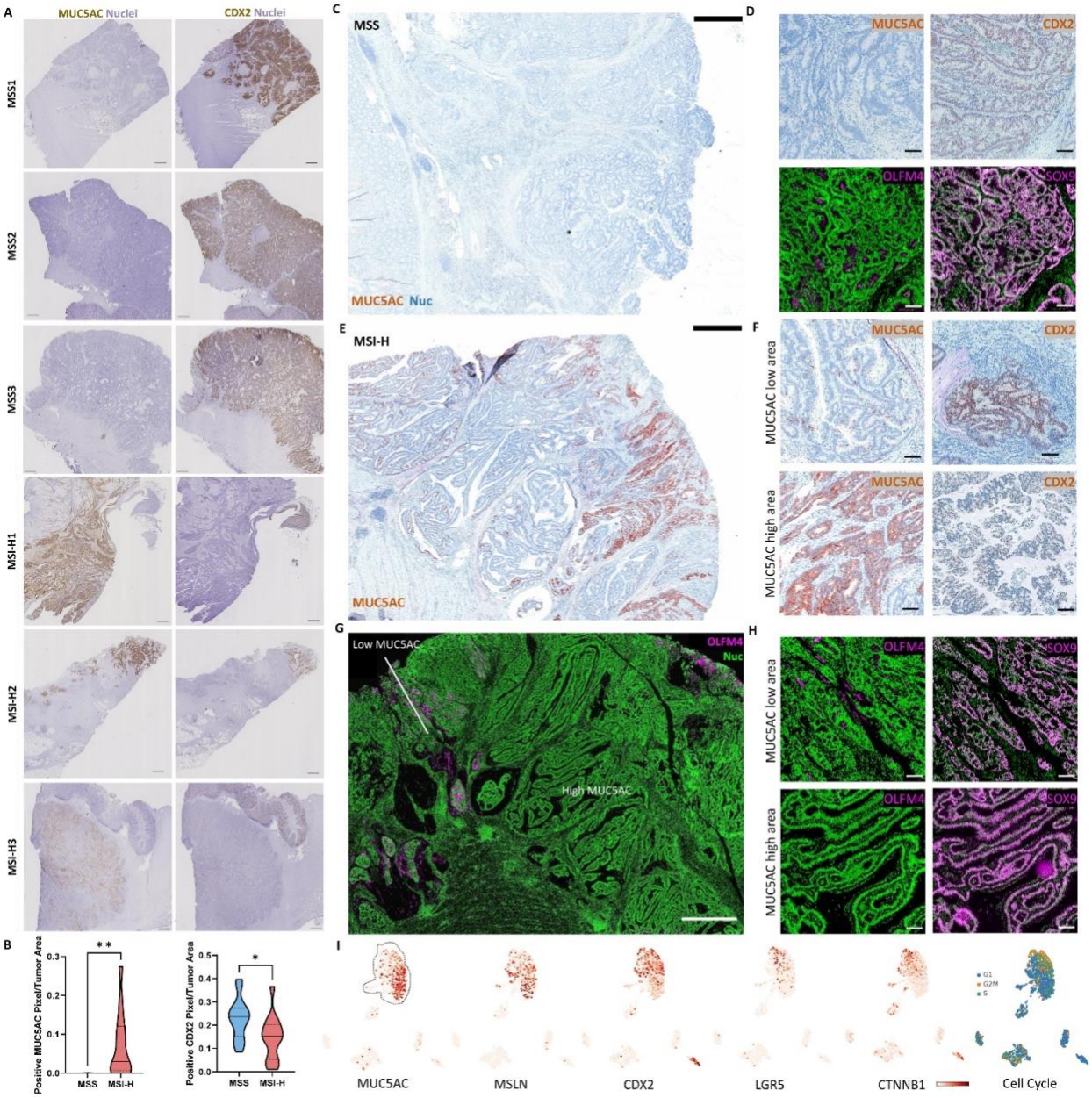

A
MUC5AC Nuclei
CDX2 Nuclei
MSS1
MSS2
MSS3
MSI-H1
MSI-H2
MSI-H3
B
Positive MUC5AC Pixel/Tumor Area
Positive CDX2 Pixel/Tumor Area
MSS
MSI-H
C
MSS
MUC5AC Nuc
D
MUC5AC
CDX2
OLFM4
SOX9
E
MSI-H
MUC5AC
F
MUC5AC low area
MUC5AC high area
G
Low MUC5AC
High MUC5AC
OLFM4
Nuc
H
I
MUC5AC
MSLN
CDX2
LGR5
CTNNB1
Cell Cycle
G1
G2M
S

Figure 24. Heterogeneity of CRCs with metaplastic and stem-like features.

(A) Representative whole slide scans for IHC of MUC5AC and CDX2 in 3 MSS and MSI-H CRCs. **(B)** Image quantification of positive pixels per tumor area in IHC scans of n=17 MSS and n=14 MSI-H CRCs, presented as violin plots with median as solid lines and quartiles as dotted lines. **(C)** Low magnification view of MUC5AC staining, and **(D)** high magnification view of a MSS CRC with various protein stains. **(E)** Low magnification view of MUC5AC staining of an MSI-H CRC. **(F)** High magnification view of a MUC5AC high and MUC5AC low area for metaplasia markers of the CRC in E. **(G,H)** Same as in E,F but staining for stem cell markers. **(I)** Count-based UMAP of scRNA-seq data of the MSI-H CRC in E overlaid with various metaplasia and stem cell markers, as well as with cell cycle signatures. Color intensity represents mean scaled and standardized Arcsinh gene expression. *p<0.05, **p<0.01

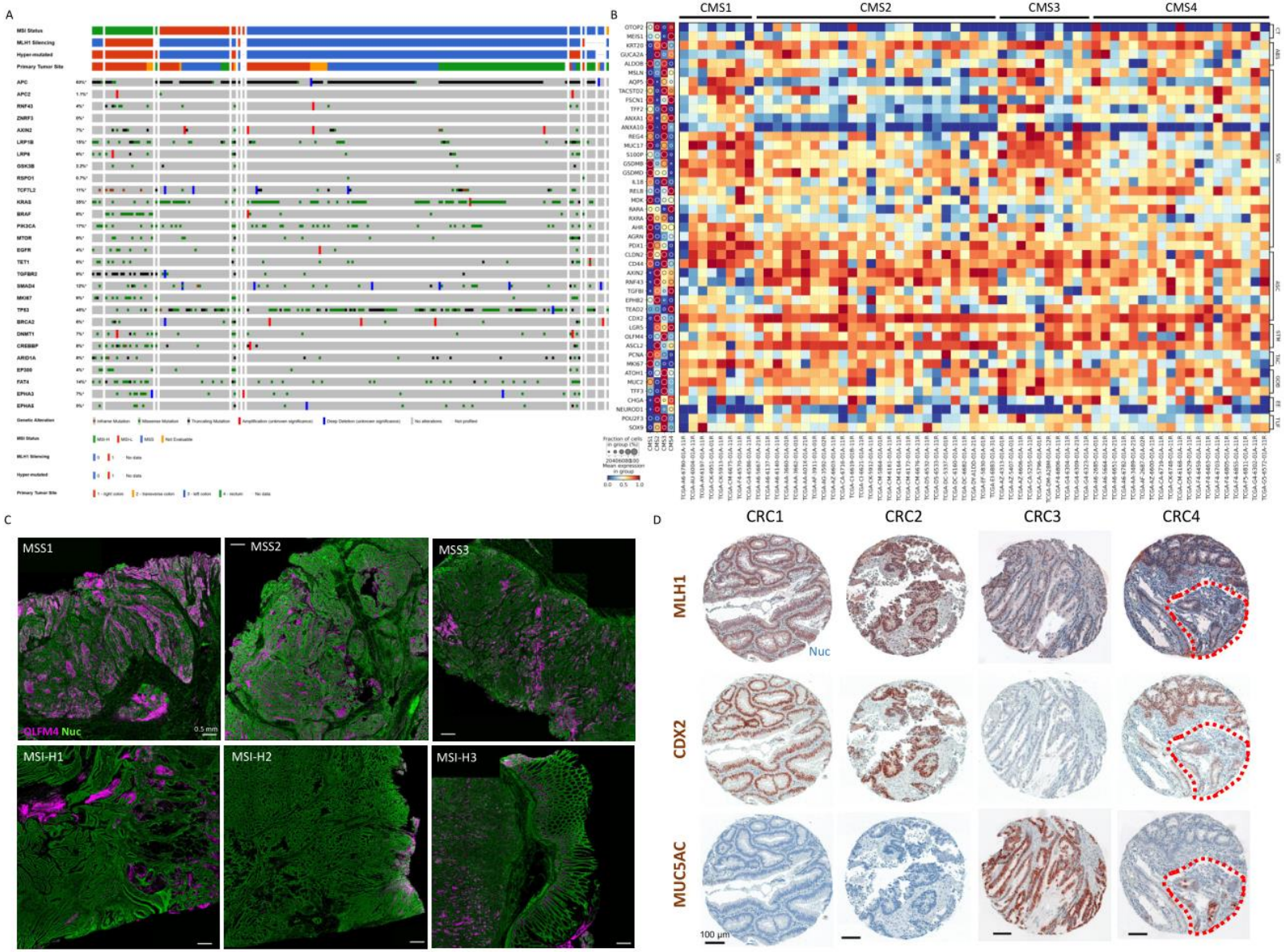

A
B
CMS1
CMS2
CMS3
CMS4
C
MSS1
MSS2
MSS3
OLFM4 Nuc
0.5 mm
MSI-H1
MSI-H2
MSI-H3
D
CRC1
CRC2
CRC3
CRC4
MLH1
Nuc
CDX2
MUC5AC
100 µm

Figure 25. Homogeneous and heterogeneous features of CRCs.

(A) Oncoplot visualizing genomic tracks for 276 assayed TCGA CRC specimens. Each row represents unsorted mutational status of key pathway genes. Each column represents an individual specimen with hypermutation status noted. **(B)** Heatmap representation of gene signatures identified by scRNA-seq of pre-cancers applied to bulk RNA-seq of TCGA CRC specimens (CMS1, n=8; CMS2, n=26; CMS3, n=10; CMS4, n=19). Inset circle on summary heatmap (left) indicates the fraction of specimens presented with gene expression and color intensity represents mean (left) or individual (right) scaled and standardized Arcsinh gene expression. **(C)** Representative whole slide scans for MxIF of OLFM4 in 3 MSS and MSI-H CRCs. **(D)** Staining for various metaplasia markers from 4 CRCs from a tissue microarray. Red outline represents the MLH1 low area in CRC 4.

We confirmed these findings in a CRC tissue microarray using MLH1 staining to infer the microsatellite status of each specimen. MLH1-high cancers (presumably MSS) had uniform CDX2, but not MUC5AC expression, regardless of whether they were well-differentiated (CRC1) or poorly differentiated (CRC2) **(Figure 25D)**. CRC3 had uniformly low MLH1 staining (presumably MSI-H) and expressed MUC5AC but not CDX2. CRC4 was heterogeneous in MLH1 staining and had areas of $CDX2^{low}$/$MUC5AC^{high}$ expression and areas with $CDX2^{high}$/$MUC5AC^{low}$ expression. These results suggest MSI-H CRCs may acquire stem cell characteristics in a background of metaplasia leading to cellular heterogeneity in tumor landscapes.

Serrated polyps associate with a $CD8^+$ T cell enriched cytotoxic microenvironment prior to developing hypermutation

Although unidentified factors likely contribute, high neoantigen load in MSI-H CRCs has been hypothesized to induce a cytotoxic microenvironment conferring responsiveness to immunotherapy [131,215]. SERs did not exhibit hypermutation in our mutational analysis **(Figure 16C; Figure 17A)**, but all MSI-H CRCs analyzed were hypermutated **(Figure 23A)**. We then sought to determine whether or not SERs have a distinct tumor microenvironmental signature preceding hypermutation. We combined analyses of the non-epithelial scRNA-seq datasets from colonic pre-cancers and cancers to identify T cells, plasma cells, myeloid cells, mast cells, fibroblasts, endothelial cells, and B cells based on differential gene expression **(Figure 26A,B; Figure 27A-C)**.

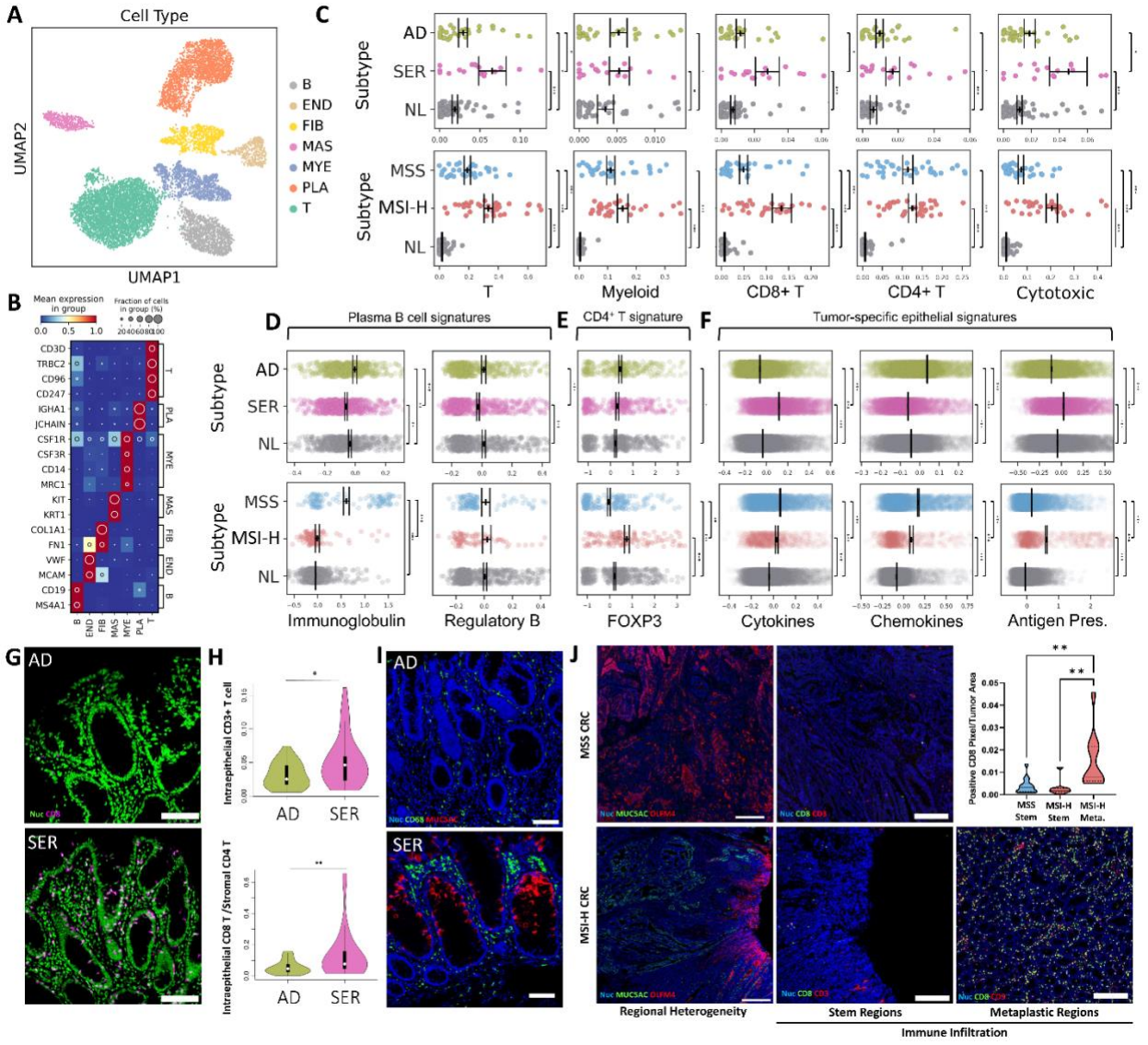


Figure 26. Immune cell characterization of colonic tumor subtypes. The immune microenvironment of different tumor subtypes.

(A) Regulon-based UMAP representation of non-epithelial cell populations color overlaid with cell type defined by Leiden clustering and marker genes. **(B)** Heatmap representation of marker genes defining each cell type in A. Inset circle indicates prevalence within defined single-cell populations and color intensity represents scaled and standardized Arcsinh gene expression. **(C)** Scatter plots of cell type representation per sample (x-axis is the proportion of cells in a sample with indicated subtype), (top) polyp and (bottom) cancer subtypes. Points represent individual specimens

derived from both VUMC and Broad cohorts. The cytotoxic subtype consists of $CD8^+$ T, NK, and γδ T cells. Error bars represent SEM of n= 28 for AD, n = 17 for SER, n = 66 for NL, n = 33 for MSS, and n = 34 for MSI-H. **(D-F)** Scatter plots of **(D)** plasma cell, **(E)** $CD4^+$ T cell**, (F)** tumor cell - specific signature scores, with each point representing a single cell, for (top) polyp and (bottom) CRC. Tumor cell data source from epithelial data in **Figures 2 and 4**. Error bars depict SEM of single cells. **(G)** Representative MxIF images of $CD8^+$ cells in ADs and SERs. **(H)** Quantification of MxIF images of intraepithelial $CD8^+$ cells, and intraepithelial $CD8^+$ to stromal $CD3^+/CD4^+$ T cells, normalized to total number of cells in each compartment. Violin plots with median as white circles and quartiles as candles for n=20 polyps per type. **(I)** Representative MxIF images of $CD68^+$ cells and $MUC5AC^+$ in ADs and SERs. **(J)** Representative MxIF scans of intratumoral heterogeneous regions within CRCs ($OLFM4^+$ - stem versus $MUC5AC^+$ - metaplasia). MSS CRC only has stem regions. Representative MxIF images of $CD8^+$ and $CD3^+$ cells within each of the stem and metaplastic regions. Inset is the quantification of CD8 positive pixels per tumor area in stem versus metaplastic regions in MxIF scans of n=15 MSS and n=10 MSI-H CRCs, presented as violin plots with median as solid lines and quartiles as dotted lines. *p<0.05, **p<0.01, ***p<0.001.

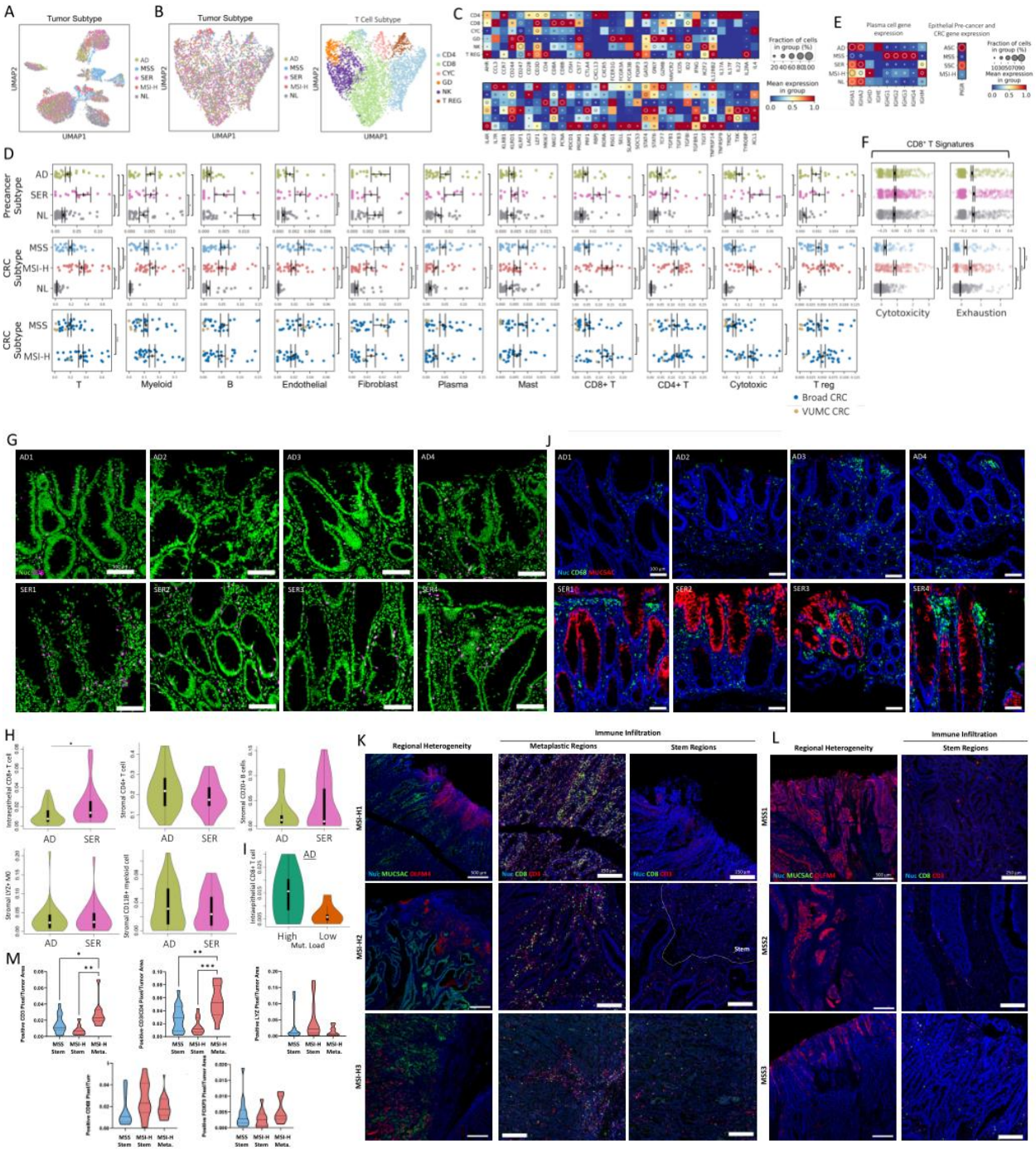


Figure 27. The immune microenvironment as related to tumor cell heterogeneity.

(A) Regulon-based UMAP representation of non-epithelial cell populations color overlaid with tissue and tumor subtype. **(B)** Feature-selected, count-based UMAP representation of T cell subtypes, color overlaid with (left) tissue and tumor subtype and (right) cell type defined by

clustering and marker gene expression. **(C)** Heatmap representation of feature-selected genes used to generate UMAP in B. Inset circle indicates prevalence within defined single-cell populations and color intensity represents scaled and standardized Arcsinh gene expression. **(D)** Comprehensive scatter plots of normalized cell type representation per tissue, (top) polyp and (middle) cancer subtypes. Points represent individual specimens. Error bars represent SEM of n= 28 for AD, n = 17 for SER, n = 66 for NL, n = 33 for MSS, and n = 34 for MSI-H. Plots in the third row colored by VUMC or Broad datasets. **(E)** Heatmap representation of marker genes defining plasma B immunoglobulin signature alongside PIGR expression in corresponding neoplastic epithelial cells. Inset circle indicates prevalence within defined single-cell populations and color intensity represents scaled and standardized Arcsinh gene expression. **(F)** Scatter plots of $CD8^+$ T cell cytotoxicity and exhaustion signatures, with each point representing a single cell, for (top) polyp and (bottom) CRC. Error bars depict SEM of single cells. **(G)** Additional representative MxIF images of $CD8^+$ cells in multiple ADs and SERs. **(H)** Quantification of MxIF images of marker positive cells in epithelial or stromal compartments, normalized to total number of cells in each compartment. Violin plots with median as white circles and quartiles as candles for n=20 polyps per type. **(I)** Quantification as in H for $CD8^+$ T cells in the epithelial compartment of TAs separated by mutational burden, for n=6 high (>1500 mutations) vs n=7 low (<1500 mutations). **(J)** Additional representative MxIF images of $CD68^+$ cells and $MUC5AC^+$ in multiple ADs and SERs. **(K,L)** Additional representative MxIF scans of intratumoral heterogeneous regions within CRCs ($OLFM4^+$ - stem versus $MUC5AC^+$ - metaplasia), and additional representative MxIF images of $CD8^+$ and $CD3^+$ cells within each of the stem and metaplastic regions of **(K)** MSI-H CRCs and **(L)** MSS CRCs with only stem regions. **(M)** Quantification of marker positive pixels per tumor area in stem versus metaplastic regions in MxIF scans of n=15 MSS and n=10 MSI-H CRCs, presented as violin plots with median as solid lines and quartiles as dotted lines. *p<0.05, **p<0.01, ***p<0.001.

Microenvironmental composition was markedly different between ADs, SERs, and normal biopsies. Most immune cell types were increased in polyps compared to normal tissue, although many were not different between subtypes **(Figure 26C; Figure 27D)**. Strikingly, $CD8^+$ T cells, NK cells, and γδT cells, the cytotoxic immune cells, were significantly increased in SERs compared to ADs **(Figure 26C).** Increases in overall T cell numbers in SERs compared to ADs were not driven by helper cells, since $CD4^+$ T cell numbers were not statistically different between the polyp types, although these cells were increased in polyps compared to normal **(Figure 26C)**. The overrepresentation of total T cells, and specifically cytotoxic T cells, was also observed in MSI-H CRCs compared to MSS CRCs, with no difference in $CD4^+$ T cells, suggesting a consistent dichotomy in the adaptive microenvironment between subtypes regardless of hypermutation.

Pre-cancer resident plasma cells express IgA, which is transported across the epithelium into the lumen by PIGR. *PIGR* was highly expressed in epithelial pre-cancer cells compared to both MSS and MSI-H cancer cells **(Figure 27E)**, suggesting the normal gut humoral response is relatively intact in polyps [216,217]. The immunoglobulin gene signature was significantly lower in SER plasma cells compared to AD plasma cells, and it was further suppressed in MSI-H compared to MSS CRCs **(Figure 26D; Figure 27E)**. MSS CRC plasma cells did not express IgA but instead expressed IgG, the major immunoglobulin subtype in blood plasma cells arising from the spleen and lymph nodes. SER plasma cells presented a diminished regulatory B signature, consistent with an active immune environment **(Figure 26D)**.

Despite differences in $CD8^+$ cytotoxic T cell abundance, gene signatures related to cytotoxicity and exhaustion did not differ between ADs and SERs **(Figure 27F)**. FOXP3 regulon activity was higher than normal colon in AD-derived $CD4^+$ T helper cells, consistent with a degree of Treg-dependent immunosuppression **(Figure 26E)**. Signatures of cytotoxicity and exhaustion were intensified in MSI-H CRC T cells compared with MSS CRC T cells, indicative of T cell dysfunction

in MSI-H CRCs **(Figure 27F)**. Differential adaptive immune cell regulation between conventional and serrated pathway tumors was also observed in tumor epithelial cells. ASCs expressed a monocyte-attracting chemokine signature, and SSCs expressed a lymphocyte-attracting cytokine signature important for establishing an adaptive immune environment **(Figure 26F)** [218,219]. An antigen processing and presentation gene signature [93,208] was significantly higher in SSCs relative to ASCs, which was also increased in MSI-H CRC cells relative to MSS CRC cells **(Figure 26F)**. These transcriptomic data illustrate how the differential regulation of adaptive immunity persists during progression to malignancy and appears independent of hypermutation status.

Using scRNA-seq results to inform multiplex imaging, we examined the geographical differences in the immune compartment between ADs and SERs. As expected, SERs had a higher number of total T cells and $CD8^+$ cytotoxic T cells, as well as a higher ratio of $CD8^+$ to $CD4^+$ T cells compared to ADs **(Figure 26G,H; Figure 27G,H)**, while other immune cell populations were not significantly different. $CD8^+$ cytotoxic T cells had a close spatial association with epithelial cells in SERs. There appeared to be more $CD8^+$ cytotoxic T cells in the stromal compartment of ADs with high mutational load versus those with a lower load; however, our analysis was underpowered to show a statistically significant difference **(Figure 27I)**. While myeloid cell abundance was not different by both scRNA-seq and imaging, the spatial distribution of $CD68^+$ macrophages was markedly different between SERs and ADs. In ADs, $CD68^+$ macrophages were distributed throughout the tumor stroma, but they were concentrated at the luminal surfaces of SERs **(Figure 26I; Figure 27J).** Macrophages in SERs appeared prominently near $MUC5AC^+$ metaplastic cells, coinciding with the surface localization of these lesions **(Figure 26I; Figure 27J)**. A similar striking distribution of $CD68^+$ macrophages was reported after fecal transplant and successful immunotherapy response [220], further supporting the influence of epithelial-microbial interactions on SER tumorigenesis and cytotoxic immune responses. MSI-H CRCs had a heterogeneous distribution of total T cells and $CD8^+$ T cells mirroring the observed tumor cell heterogeneity. There

was a significant enrichment of $CD8^+$ T cells in $MUC5AC^+$ metaplastic regions and reduced numbers in $OLFM4^+$ stem cell-like regions **(Figure 26J; Figure 27K-M)**. In contrast, MSS CRCs had fewer T cells throughout the tumors, which were homogeneously composed of $OLFM4^+$ stem-like cells **(Figure 26J; Figure 27K-M)**. These results strengthen the association between the metaplastic origin of SERs and the cytotoxic immune microenvironment, and implicate immune suppression as tumor cells gain stem properties.

Tumor cell differentiation status dictates the adaptive immune microenvironment

We next used mouse models to determine if the cytotoxic response in serrated tumorigenesis is intrinsic to tumor cell state. Our human SER data and a recent mouse model [221] demonstrate that a cytotoxic immune environment is established prior to the onset of hypermutation and microsatellite instability. We thus investigated the underpinnings of the induction of cytotoxic immunity using genetically engineered mice that model the earliest events of tumorigenesis. The *Lrig1*$^{CreERT2/+}$;*Apc*$^{2lox14/+}$ is a well-established model of conventional pathway tumorigenesis, resulting in development of adenomatous tumors in the distal colon [111]. Driving a *Braf* activating mutation (*Lrig1*$^{CreERT2/+}$;*Braf*$^{LSL\text{-}V600E/+}$) did not result in macroscopic tumors, but induced villiform metaplasias in the proximal colon **(Figure 28A)**. *Apc* mutant tumors had elevated cytoplasmic staining of β-catenin and a reduced number of $CD8^+$ T cells compared to control normal colon, consistent with human ADs and MSS CRCs **(Figure 28B,C)**. In contrast, *Braf* mutant lesions were associated with increased $CD8^+$ T cell infiltration **(Figure 28B,C)**. Together, these results suggest that the different immune responses in tumors from conventional and serrated pathways can each be modeled via a single driver mutation. Strikingly, $CD8^+$ T cell infiltration was only observed in the differentiated cell compartment of villiform metaplasias and not in the mutant crypts, signifying that *Braf* mutant differentiated cells, but not stem cells, drive the cytotoxic microenvironment **(Figure 28C)**. Similar results were observed in a parallel *Kras*-activating mouse model (*Lrig1*$^{CreERT2/+}$;*Kras*$^{LSL\text{-}G12D/+}$) **(Figure 29A-C)**.

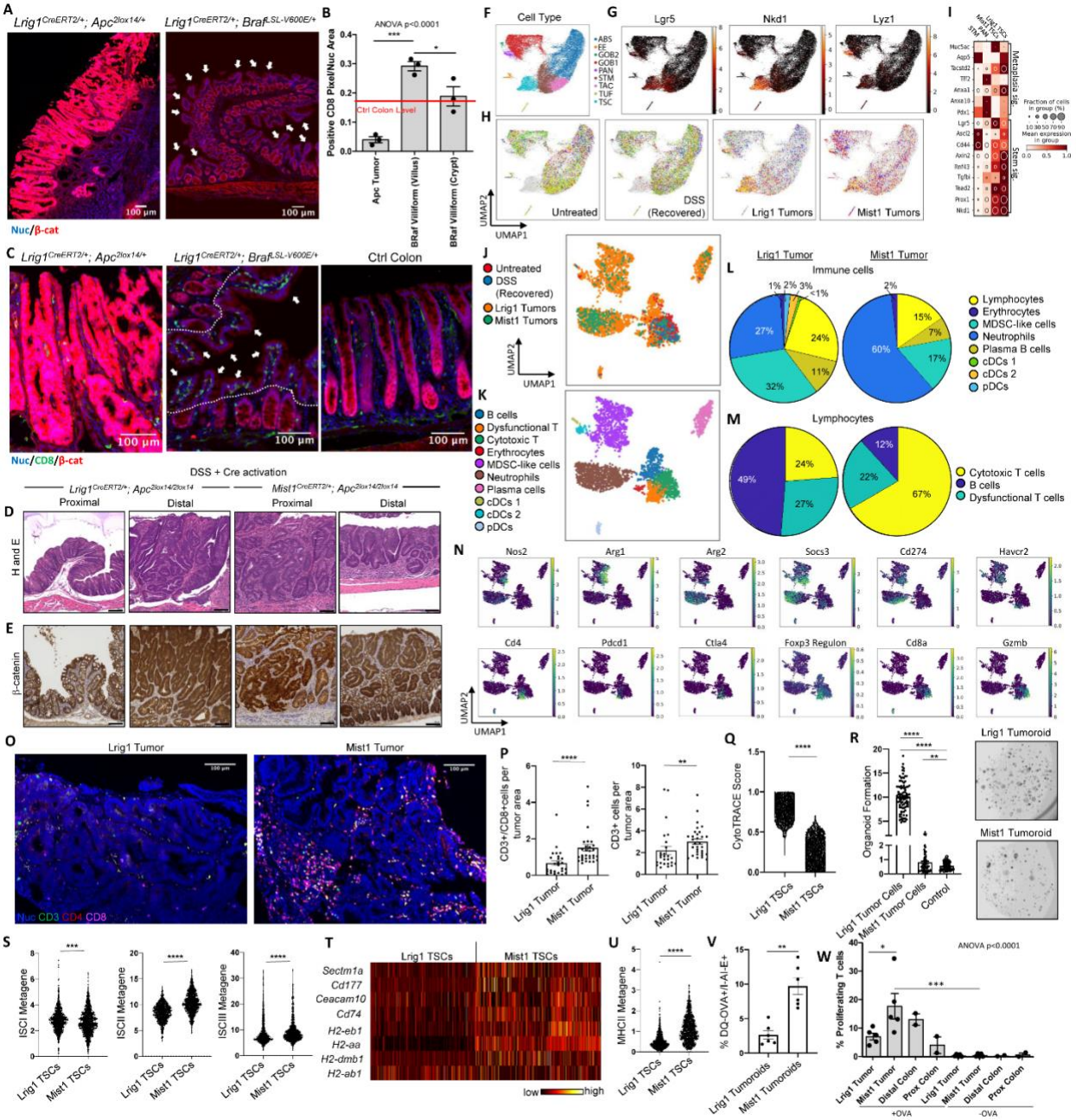

A
Lrig1CreERT2/+; Apc2lox14/+
Lrig1CreERT2/+; BrafLSL-V600E/+
100 μm
Nuc/β-cat
B
ANOVA p<0.0001
Positive CD8 Pixel/Nuc Area
Ctrl Colon Level
Apc Tumor
BRaf Villiform (Villus)
BRaf Villiform (Crypt)
C
Lrig1CreERT2/+; Apc2lox14/+
Lrig1CreERT2/+; BrafLSL-V600E/+
Ctrl Colon
Nuc/CD8/β-cat
DSS + Cre activation
Lrig1CreERT2/+; Apc2lox14/2lox14
Mist1CreERT2/+; Apc2lox14/2lox14
Proximal
Distal
H and E
β-catenin
F
Cell Type
ABS
EE
GOB2
GOB1
PAN
STM
TAC
TUF
TSC
G
Lgr5
Nkd1
Lyz1
H
Untreated
DSS (Recovered)
Lrig1 Tumors
Mist1 Tumors
UMAP1
UMAP2
I
Lrig1 TSCs
Mist1 TSCs
STM
Metabolism sig.
Stem sig.
Fraction of cells in group (%)
Mean expression in group
J
Untreated
DSS (Recovered)
Lrig1 Tumors
Mist1 Tumors
K
B cells
Dysfunctional T
Cytotoxic T
Erythrocytes
MDSC-like cells
Neutrophils
Plasma cells
cDCs 1
cDCs 2
pDCs
L
Lrig1 Tumor
Mist1 Tumor
Immune cells
1% 2% 3% <1%
27%
24%
11%
32%
2%
15%
7%
60%
17%
Lymphocytes
Erythrocytes
MDSC-like cells
Neutrophils
Plasma B cells
cDCs 1
cDCs 2
pDCs
M
Lymphocytes
12%
24%
49%
27%
22%
67%
Cytotoxic T cells
B cells
Dysfunctional T cells
N
Nos2
Arg1
Arg2
Socs3
Cd274
Havcr2
Cd4
Pdcd1
Ctla4
Foxp3 Regulon
Cd8a
Gzmb
O
Lrig1 Tumor
Mist1 Tumor
Nuc CD3 CD4 CD8
P
CD3+/CD8+cells per tumor area
CD3+ cells per tumor area
Lrig1 Tumor
Mist1 Tumor
Q
CytoTRACE Score
Lrig1 TSCs
Mist1 TSCs
R
Organoid Formation
Lrig1 Tumor Cells
Mist1 Tumor Cells
Control
Lrig1 Tumoroid
Mist1 Tumoroid
S
ISCI Metagene
ISCII Metagene
ISCIII Metagene
Lrig1 TSCs
Mist1 TSCs
T
Lrig1 TSCs
Mist1 TSCs
Sectm1a
Cd177
Ceacam10
Cd74
H2-eb1
H2-aa
H2-dmb1
H2-ab1
low
high
U
MHCII Metagene
Lrig1 TSCs
Mist1 TSCs
V
% DQ-OVA+/I-A/I-E+
Lrig1 Tumoroids
Mist1 Tumoroids
W
ANOVA p<0.0001
% Proliferating T cells
Lrig1 Tumor
Mist1 Tumor
Distal Colon
Prox Colon
+OVA
-OVA

Figure 28. Functional validation of the tumor cell differentiation status and the effects on cytotoxic immunity.

(A) Representative macroscopic IF images depicting a *Lrig1*$^{CreERT2/+}$*; Apc*$^{2lox14/+}$ tumor and *Lrig1*$^{CreERT2/+}$*; Braf*$^{LSL\text{-}V600E/+}$ proximal villiform metaplasia (white arrows). **(B)** Quantification of CD8 positive pixels per nuclear area in *Apc*-driven tumor, and the villus and crypt compartment of *Braf*-driven villiform metaplasias. Red line denotes the mean level detected in adjacent normal colon in *Braf* mice. Error bars represent SEM from n=3 animals per group. **(C)** IF images of CD8$^+$ T cells in an *Apc*-driven tumor overexpressing β-catenin, *Braf*-driven villiform metaplasias (white arrows), and control colon. Dotted line demarcates border between villus and crypt compartments. **(D,E)** Representative H&E **(D)** and β-catenin IHC **(E)** of colonic tissues and tumors from induced *Lrig1*$^{CreERT2/+}$*; Apc*$^{2lox14/2lox14}$ mice (n=4 animals; advanced dysplasia) and induced *Mist1*$^{CreERT2/+}$*; Apc*$^{2lox14/2lox14}$ mice (n=4; low grade dysplasia) 28 days after DSS. **(F-H)** Combined UMAP embedding of epithelial scRNA-seq data generated from mouse colonic tissues (n=3 or 4 per condition), with overlays indicating **(F)** Leiden clustering labeled by cell populations (ABS-absorptive, EE-enteroendocrine, GOB2-goblet 2, GOB1-goblet 1, PAN-Paneth, STM-stem TAC-transit-amplifying, TUF-tuft, TSC-tumor specific cell), **(G)** specific gene overlays with color intensity representing scaled and standardized Arcsinh gene expression, **(H)** distribution of different biological replicates (mice) under specific conditions. **(I)** Heatmap representation of gene sets defining a human metaplastic or stem cell signature in specific cell populations, including Mist1 and Lrig1 TSCs. Inset circle indicates prevalence within defined single-cell populations and color intensity represents scaled and standardized Arcsinh gene expression. **(J,K)** Combined UMAP embedding of scRNA-seq data of immune cells generated from colonic tissues from C, with overlays indicating **(J)** different conditions and **(K)** Leiden clustering labeled by cell populations. **(L,M)** Quantification of **(L)** general immune cell types and **(M)** specific lymphocyte populations in Lrig1 (left) and Mist1 (right) tumors. **(N)** UMAP overlays of specific gene expression delineating immunosuppression or cytotoxicity in myeloid and lymphoid cell lineages. Color

intensity represents scaled and standardized Arcsinh gene expression. **(O)** Representative MxIF images of Lrig1 (left) and Mist1 (right) tumors with markers delineating T cells. **(P)** Normalized quantification of (left) $CD3^+/CD8^+$ and (right) $CD3^+$ cells per tumor area in Lrig1 and Mist1 tumors. Each dot represents a field of view. Error bars represent SEM from n=3 animals per group. **(Q)** CytoTRACE score distribution, a predictor of stemness, for Lrig1 and Mist1 TSCs calculated from scRNA-seq. **(R)** Normalized organoid formation efficiency of single cells isolated from Lrig1 tumors, Mist1 tumors, and control colons. Each dot represents data from a well with representative images shown in insets. Error bars represent SEM from n=4 animals per tumor, 2 for control. **(S)** Normalized metagene signature expression of ISCI, ISCII, and ISCIII for Lrig1 and Mist1 TSCs derived from scRNA-seq. **(T)** Heatmap of individual antigen presentation, scaled and standardized Arcsinh gene expression at single-cell level. **(U)** Normalized MHCII metagene signature expression for Lrig1 and Mist1 TSCs. **(V)** Quantification of flow cytometry plots of $DQ\text{-}OVA^+/I\text{-}AI\text{-}E^+$ epithelial cells comparing antigen processing and presentation abilities between Lrig1- and Mist1- tumoroids. Error bars represent SEM from n=6 animals per condition. **(W)** Percentage of proliferating T cells determined by CellTrace Violet assay when cocultured with organoids derived from colonic tumors or normal tissues (+DSS) treated with or without x 100 ug/ml OVA peptide. Error bars represent SEM of organoids from n=5 mice for tumors and 2 for normal. *p<0.05, **p<0.01, ***p<0.001, ****p<0.0001.

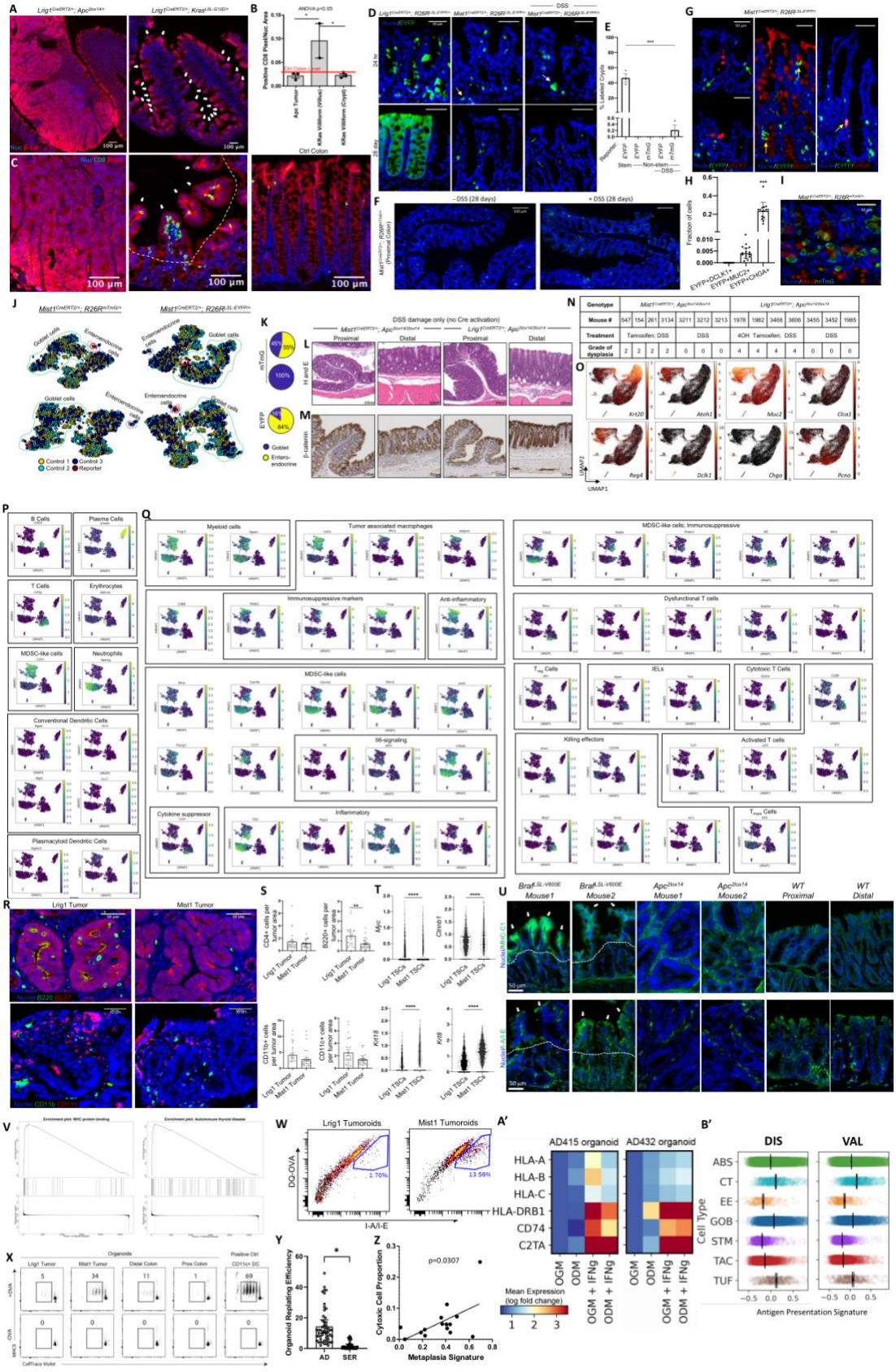


Figure 29. Stem versus non-stem cell characteristics of tumors and the tumor immune environment.

(A) Representative macroscopic IF images depicting a *Lrig1*$^{CreERT2/+}$*; Apc*$^{2lox14/+}$ tumor and *Lrig1*$^{CreERT2/+}$*; Kras*$^{LSL\text{-}G12D/+}$ proximal villiform metaplasia (white arrows). **(B)** Quantification of CD8 positive pixels per nuclear area in *Apc*-driven tumor, and the villus and crypt compartment of *Kras*-driven villiform metaplasias. Red line denotes the mean level detected in adjacent normal colon in *Kras* mice. Error bars represent SEM from n=3 animals per group. **(C)** IF images of CD8+ T cells in an *Apc*-driven tumor overexpressing β-catenin, *Kras*-driven villiform metaplasias (white arrows), and control colon. Dotted line demarcates border between villus and crypt compartments. **(D)** Representative IF images of short term and long term (rows: 24 hours and 28 days, respectively) lineage tracing in the colon using *Lrig1*$^{CreERT2/+}$*; R26R*$^{LSL\text{-}EYFP/+}$, *Mist1*$^{CreERT2/+}$*; R26R*$^{LSL\text{-}EYFP/+}$, and Mist1$^{CreERT2/+}$*; R26R*$^{LSL\text{-}EYFP/+}$ with DSS mouse models (columns). **(E)** Quantification of lineage traced glands at 28 days at homeostasis or with DSS-damage from stem (Lrig1) and non-stem (Mist1) cells using different reporters. Error bars represent SEM from n=3 animals for each condition. **(F)** Representative IF images of long term (28 days) lineage tracing in the proximal colon of Mist1$^{CreERT2/+}$*; R26R*$^{LSL\text{-}mTmG/+}$ with or without DSS. **(G)** Representative IF images of reporter co-expression with cell type specific markers, DCLK1 for tuft cells, MUC2 for goblet cells, and CHGA for enteroendocrine cells, after 24 hours of lineage tracing in Mist1$^{CreERT2/+}$*; R26R*$^{LSL\text{-}EYFP/+}$ mouse colons. Arrows point to single positive (green/red) or double positive (yellow) cells. **(H)** Quantification of double positive cells compared to single positive cells for each marker. Each dot represents a field of view. Error bars represent SEM from n=3 animals. **(I)** Representative IF image of reporter co-expression with MUC2 in *Mist1*$^{CreERT2/+}$*; R26R*$^{mTmG/+}$ 10 days after initiation of lineage tracing at homeostasis. **(J)** Combined t-SNE embedding of bulk RNA sequenced *Mist1* reporter (mTmG – left or EYFP – right)-expressing cells (solid circle in red) with n=3 reference murine colonic scRNA-seq dataset (colored dots from different mice). Goblet (green) and enteroendocrine (magenta) cell populations are delineated by dotted lines. **(K)**

Quantification of *Mist1* reporter (EYFP or mTmG)-expressing cells co-embedding with different cell types using boot-strapped t-SNE runs in J. n=3 independent experiments. **(L,M)** Representative H&E **(L)** and β-catenin IHC **(M)** of control (uninduced) normal colonic tissues as in Figure 28D,E. **(N)** Histological scoring of tissues of multiple biological replicates of Mist1- and Lrig1- tumors along with uninduced controls. 0: none; 1: unicrypt; 2: low; 3: high, 4: advanced. **(O)** Combined UMAP embedding as in Figure 28F with overlays of key marker gene expression delineating epithelial cell types and states. Color intensity represents scaled and standardized Arcsinh gene expression. **(P,Q)** Combined UMAP embedding as in Figure 28J with overlay of **(P)** marker gene expression delineating immune cell types and **(Q)** expression of genes involved in immunosuppression, cytotoxicity, and effector function. Color intensity represents scaled and standardized Arcsinh gene expression. **(R)** Representative MxIF images for tumor infiltrating immune cells in Lrig1- versus Mist1- tumors. **(S)** Quantification of immune cells from MxIF. Error bars represent SEM from n=3 animals. Dots represent fields of view. **(T)** Gene expression of *Myc, Ctnnb1, Krt18*, and *Krt8* for Lrig1 and Mist1 TSCs from scRNA-seq. Library size-normalized and Arcsinh-scaled. **(U)** Selected GSEA enrichment plots generated using KEGG and Gene Ontology of Mist1 TSCs against Lrig1 TSCs. **(V)** Representative flow cytometry plots of DQ-OVA^{+}/I-AI-E^{+} epithelial cells comparing antigen processing and presentation abilities between Lrig1- and Mist1-tumoroids. **(W)** Representative flow cytometry plots of CellTrace Violet peaks that depicts proliferating T cells when cocultured with organoids derived from colonic tumors or normal tissues (+DSS) treated with or without x 100 ug/ml OVA peptide. Positive control is dendritic cells used for antigen presentation. Gates denote proliferating cells plotted in Figure 28W. **(X)** Organoid replating efficiency as presented as number of organoids formed from 1000 single cells isolated from human ADs and SERs. Each dot represents data from a well. Error bars represent SEM from n=7 ADs and 6 SERs. *< p<0.05, **p<0.01, ***p<0.001, ****p<0.0001.

To determine how a differentiated cell versus stem cell state influences the immune microenvironment, we normalized the genetic event by driving the same *Apc* mutation from stem (*Lrig1*^{CreERT2}) versus non-stem (*Mist1*^{CreERT2}) cells. While *Lrig1*$^{+}$ cells are bona fide stem cells (Powell et al., 2012), lineage tracing studies during both homeostasis and DSS damage showed *Mist1*$^{+}$ cells are non-stem cells in the proximal colon **(Figure 29D,E)**. No lineage tracing was observed from *Mist1*$^{+}$ cells in the colon under any condition using a standard YFP reporter **(Figure 29D,E)**. A baseline level of lineage tracing was observed only in the distal colon of a more sensitive mT/mG reporter, and not in the proximal colon, even with DSS damage **(Figure 29F)**. Using immunostaining and transcriptomics, we determined *Mist1*$^{+}$ cells represent a subset of differentiated cells outside of the colonic crypt base (goblet/enteroendocrine) **(Figure 29G-K)**.

Importantly, *Mist1*$^{+}$ cells initiated colonic tumors (abbreviated as Mist1 tumors) with biallelic recombination of *Apc* (*Mist1*$^{CreERT2/+}$*; Apc*$^{2lox14/2lox14}$) followed by 2.5% DSS damage, representing a non-stem-driven tumor model. At most one or two Mist1 tumors developed per mouse in the proximal versus distal colon by a 7:1 ratio **(Figure 28D,E; Figure 29L,M)**, which differs from the distal colon predominance of tumors in the *Lrig1*$^{CreERT2/+}$*; Apc*$^{2lox14/+}$ model (Powell et al., 2012). We developed a stem cell-driven tumor model (abbreviated as Lrig1 tumors) for direct comparison with the *Apc* mutation using *Lrig1*$^{CreERT2/+}$*; Apc*$^{2lox14/2lox14}$ mice and focal Cre activation, followed by DSS **(Figure 28D,E; Figure 29L,M)**. Blinded histological assessment revealed Lrig1 tumors were high-grade dysplastic tumors, but Mist1 tumors were low-grade **(Figure 29N)**. To decipher the molecular landscape of the two tumor types, we performed scRNA-seq on harvested tumor tissues along with controls (untreated colon and after DSS recovery) (**Figure 28F,G; Figure 29O**) and identified cells specific to tumors but not controls **(Figure 28H)**, similar to our human study. Abnormal Paneth cells were only present in tumors and not in the normal colon **(Figure 28F,G)**. Due to a common WNT-driven mutational process, tumor specific cells (TSCs) from both tumor types formed an *Lgr5*-overexpressing cell population without a metaplastic gene signature

(Figure 28G-I), similar to ASCs. Moreover, both tumor types exhibited elevated cytoplasmic and nuclear staining of β-catenin staining reflecting WNT activation **(Figure 28E; FigureS7M)**.

While the mutational processes between the tumor types were the same, we identified marked differences in the immune microenvironments **(Figure 28J)**. Mist1 tumors, similar to SERs, harbored higher proportions of $CD8^+$ T cells, which clustered with intraepithelial lymphocytes (IEL) from control colons and expressed IEL markers *Itgae* and *Trdc* [222,223] **(Figure 28J-M; Figure 29P).** These cells expressed markers of active cytotoxicity and killing effectors, which are summarized in **Supplemental Table S6 (Figure 28N; Figure 29Q)**. Lrig1 tumors possessed a distinct population of dysfunctional $CD4^+$ T cells that may have transitioned into anergy or exhaustion and ultimately contributes to immunosuppression **(Figure 28J-M; Figure 29P)** [224–226]. These cells expressed immunosuppressive markers, the most prominent being *Pdcd1* (PD1) [227], *Ctla4* [228], *Prdm1* [229], and *Havcr2* (TIM3) [230], as well as genes of the *Foxp3* regulon, implicating dysfunctional T cells exhibiting regulatory characteristics (**Figure 28N; Figure 29Q)**. Strikingly, Lrig1 tumors, but not Mist1 tumors, had a large infiltration of myeloid cells including tumor-associated macrophages and myeloid derived suppressive-like cells, and distinct neutrophils that expressed *Cd274* (PDL1) **(Figure7J-N; Figure 29P,Q)**. Our scRNA-seq results informed multiplex imaging experiments that showed a significantly higher number of tumor-infiltrating $CD8^+$ T cells but not $CD4^+$ T cells in Mist1 tumors compared to Lrig1 tumors **(Figure 28O,P)**. Lrig1 tumors had a greater infiltration of B cells consistent with scRNA-seq data **(Figure 29R,S)**. In separate mouse models with identical *Apc* mutations, we show tumors originating from differentiated cells promote a cytotoxic immune microenvironment while tumors driven by stem cells associate with a suppressive immune microenvironment.

We then aimed to identify whether epithelial cell-intrinsic stemness contributes to immune microenvironmental differences. Similar to human studies, we applied CytoTRACE to score stem

cell potential of the two TSC types. Lrig1 TSCs had significantly higher inferred stem potential than Mist1 TSCs **(Figure 28Q)**, which was verified by the expression of specific stem and differentiated cell genes **(Figure 29T)**. This result was validated by organoid experiments demonstrating Lrig1 tumor cells were significantly more successful in forming organoids than Mist1 tumor cells **(Figure 28R)**, which confirmed how stemness is a function of cellular origin. Gleaning from previous work defining a gradient of stemness (ISCI > ISCII > ISCIII) in normal intestinal stem cells associated with immune cell interactions [196], we found Lrig1 TSCs exhibited a higher ISCI score while Mist1 TSCs exhibited higher ISCII and ISCIII scores **(Figure 28S)**; ISCII and ISCIII were found to exhibit higher antigen processing and presentation abilities. Consistent with this finding, Mist1 TSCs had increased expression of antigen presentation machinery, both at the single-cell level and signature level **(Figure 28T,U)**. *Lrig1*$^{CreERT2/+}$*;Braf*$^{LSL\text{-}V600E/+}$ villiform metaplasias also exhibited increased epithelial expression of antigen presentation machinery compared to *Lrig1*$^{CreERT2/+}$*;Apc*$^{2lox14/+}$ tumors, but only in the differentiated and not in the stem cell compartment **(Figure 29U)**, while stromal expression was not different. GSEA demonstrated Mist1 TSCs were significantly enriched for genes associated with immune-mediated processes, with antigen presentation being the most significant **(Figure 29V)**. These results demonstrate how the degree of stemness within neoplastic compartments, as dictated by cellular origins, is linked to the tumor immune microenvironment.

To validate expression of antigen presentation machinery actually reflects function, we assayed for antigen processing and presentation in Lrig1- and Mist1- tumor-derived tumoroids using the class 2 antigen ovalbumin (OVA). Mist1 tumoroids were shown to process and present more antigen than Lrig1 tumoroids, denoted by green fluorescence from endocytosis and proteolysis of DQ-OVA, coupled to I-A/I-E staining indicating surface antigen presentation **(Figure 28V; Figure 29W)**. In support of this observation, Mist1 tumoroids had an increased ability to stimulate T cell proliferation upon presentation of OVA peptide compared to Lrig1 tumoroids **(Figure 28W; Figure**

29X); suppression of this effect was observed in Lrig1 tumoroids compared to normal distal colonoids collected under the same condition. Human tumor organoid assays revealed a decrease in stem capacity alongside an increased antigen presentation gene signature in human SERs compared to ADs **(Figure 29X; Figure 26F)**. Within each tumor, cytotoxic cell infiltration positively correlated with metaplastic signatures in SERs **(Figure 29Z)**. Differentiation media, IFNγ (representative of type 1 immune environment found in SERs), or the two combined were used to induce human AD tumor organoids. All three conditions increased expression of antigen presentation machinery, although the effect of IFNγ was greater **(Figure 29A')**. These results are consistent with the expression of antigen presentation machinery in stem and differentiated cell types of the human colon **(Figure 29B').** Together, our data implicate how differentiation and stemness influence antigen presentation ability, which may partly underlie the differential stimulation of a cytotoxic immune response.

Methods

Mouse Models

All animal experiments were performed under protocols approved by the Vanderbilt University Animal Care and Use Committee and in accordance with NIH guidelines. Mice were 8 weeks old at the start of experiments and were humanely euthanized at the end of experiments according to approved guidelines. Animal weights were recorded at initiation of experiment and at the time of euthanasia. All animals used in this study were predominantly of the C57BL/6J background and both sexes were used. Littermate controls were used for experiments when possible. All animals were housed 2 to 5 per cage in a controlled environment in standard bedding with a standard 12-hour daylight cycle, cessation of light at 6 PM, and free access to standard chow diet and water. Experiments were conducted during the light cycle, excluding continuous dietary interventions.

Human Organoids

Polyps were dissociated and washed as described in the **COLON MAP scRNA-seq, Encapsulation and Library Generation** section. After dissociation, cells were washed 3 times with PBS containing 10 μM ROCK inhibitor (STEMCELL Technologies) and pelleted by quick-pulse centrifugation for 7 seconds. Human organoid models were generated from COLON MAP individuals of both sexes (70% female, 30% male). Polyp-derived cells were grown with Human IntestiCult organoid growth media (STEMCELL Technologies) supplemented with 10 μM Y-27632, 10 nM Gastrin I (Sigma-Aldrich), 1 mM N-acetyl-L-cysteine (Sigma-Aldrich), 500 nM A83-01 (Tocris), 50 ng/mL FGF-2 (Thermo Fisher), 100 ng/mL IGF-1 (BioLegend), 100 μg/mL Primocin (InvivoGen), and Matrigel (Corning) in a 3:1 ratio of Matrigel to media. Media was replaced every 2-3 days, and passaging was performed by dissociating the organoids in TrypLE Express (Thermo Fisher) with 10 μM Y-27632 for 15 minutes at 37 °C while shaking and triturating.

Mouse Organoids

Mouse organoids were generated from the same pool of mice used in mouse model experiments, with both sexes being used. Mouse tumors were dissociated using TrypLE Express, and cell pellets were resuspended in Matrigel and seeded in 25 μL droplets in a 24-well or 12-well plate. Once solidified, samples were incubated in 1 mL Mouse IntestiCult culture medium (STEMCELL Technologies) with 100 μg/mL Primocin for 5 days. Fresh media was replaced on day 3. Passaging was performed similarly to human organoids.

COLON MAP and CHTN TMA MxIHC

MxIHC was performed by iterative antibody staining and chromogen removal based on the protocol in [231]. Chromogen was removed between sequential rounds through sequential alcohol baths, and antibody was stripped by high temperature (95°C for 15 minutes). Single antibody stains using 3,3'-Diaminobenzidine were performed using standard protocols.

COLON MAP and CHTN TMA MxIF

Cyclical antibody staining, detection, and dye inactivation was performed as described previously by [232]. Briefly, fluorescence imaging was performed on a GE IN Cell Analyzer 2500 using the Cell DIVE platform. Images were acquired at x200 magnification with exposure times determined for each antibody. For each round of staining, DAPI images were aligned using rigid transformations to the first imaging round. The registered images were corrected for uneven illumination and autofluorescence was removed for each channel.

COLON MAP Pre-cancer Organoid Replating Efficiency Assay

COLON MAP samples that successfully formed organoids were dissociated and counted using Bio-RAD TC20 automated cell counter and plated at 1,000 cells/well in 5 µL Matrigel domes in a 96-well plate. Organoids were imaged and counted using an inverted microscope (Fisherbrand) after 8 days in culture. Patient IDs were matched to histopathology results after compilation and tabulation of results. GraphPad Prism 9 was used for plotting and statistical analysis using unpaired t-tests.

COLON MAP Pre-cancer Organoid Differentiation Assay

COLON MAP organoids were cultured in appropriately supplemented Human IntestiCult organoid growth media (OGM) for 3 days. They either remained in OGM for control or switched to supplemented Human IntestiCult organoid differentiation media (ODM) (STEMCELL Technologies) for 3 more days. For IFN gamma treatment, human recombinant IFN-gamma (Biolegend) was added to each media condition at 100 ng/mL for 24 hours prior to harvesting.

Murine Lineage Tracing

For homeostatic lineage tracing studies, *Lrig1*$^{CreERT2/+}$;*Rosa26*$^{LSL\text{-}EYFP/+}$ mice were injected intraperitoneally (i.p.) for 3 consecutive days with 2.5 mg tamoxifen (Sigma-Aldrich; T5648) in

corn oil, while *Mist1*$^{CreERT2/+}$;*Rosa26*$^{LSL\text{-}EYF/+P}$ were injected i.p. for 3 consecutive days with 5 mg tamoxifen. Mice were euthanized 24 h, 10 days, and 28 days later. For damage-induced lineage tracing, *Mist1*$^{CreERT2/+}$;*Rosa26*$^{LSL\text{-}EYFP/+}$ and *Mist1*$^{CreERT2/+}$;*Rosa26*$^{mT/mG/+}$ mice were injected i.p. for 3 consecutive days with 5 mg tamoxifen, and were then administered 2.5% DSS (TdB Consultancy; Batches DB001-37, DB001-42) in drinking water for the following 6 days. After cessation of DSS, mice were euthanized 24 h, and 28 days later.

Murine Induction of Recombination Using Different Promoters

To recombine genes, *Lrig1*$^{CreERT2/+}$;*Braf*$^{LSL\text{-}V600E/+}$ and *Lrig1*$^{CreERT2/+}$;*Apc*$^{2lox14/+}$ mice were induced and have their tissues harvested using established protocols (Kondo et al., 2020; Powell et al., 2012)**.** Tissues were harvested from these mice approximately 12 weeks after induction of recombination. *Lrig1*$^{CreERT2/+}$;*Kras*$^{LSL\text{-}G12D/+}$ mice were anesthetized and induced with 100 µL of 10 mg/mL 4-hydroxytamoxifen (Sigma-Aldrich) in ethanol delivered with an enema using a gavage feeding needle, and tissues were harvested around 8 weeks later.

For generating tumors, *Mist1*$^{CreERT2/+}$;*Apc*$^{2lox14/2lox14}$ were injected intraperitoneally for 3 consecutive days with 5 mg tamoxifen in corn oil. They were administered 2.5% DSS in drinking water for the following 6 days, followed by a 9-day rest period, and a second round of DSS. *Lrig1*$^{CreERT2/+}$;*Apc*$^{2lox14/2lox14}$ were injected with 0.01mM 4-hydroxytamoxifen through colonoscopy-guided orthotopic injections into the mucosal lining of the distal colon [233], and were administered 2.5% DSS in drinking water for the following 6 days. Control mice received PBS injections followed by DSS. Mice were euthanized approximately 28 days following Cre induction.

Murine Immunofluorescence and Histological Imaging

Upon euthanasia of an animal, colonic tissue was removed, washed with 1X DPBS, spread longitudinally onto Whatman filter paper and fixed in 4% PFA (Thermo Scientific) overnight. Fixed

tissues were washed with 1X DPBS, swiss-rolled, and stored in 70% EtOH until processing and paraffin embedding. Tissues were sectioned at 5 mm thick onto glass slides. Slides were processed for deparaffinization, rehydration, and antigen retrieval using citrate buffer (pH 6.0; Dako) for 20 minutes in a pressure cooker at 105°C followed by a 20-minute bench cool down. Endogenous background signal was reduced by incubating slides in 1% H_2O_2 (Sigma-Aldrich) for 10 minutes, before blocking for 30 minutes in 2.5% Normal Donkey Serum in 1X DPBS prior to antibody staining. Primary antibodies against selected markers were incubated on the slides in a humidity chamber overnight, followed by three washes in PBS, and 1 hour incubation in Hoechst 33342 (Invitrogen), and compatible secondaries (1:500) conjugated to Invitrogen AlexaFluor-488 (AF-488) or Invitrogen AF-647. Slides were washed in 1X DPBS, mounted in Prolong Gold (Invitrogen) and imaged using a Zeiss Axio Imager M2 microscope with Axiovision digital imaging system (Zeiss; Jena GmBH). Multiplexed imaging using an immune cell-based antibody panel was performed by using a multiplex iterative staining and fluorescence-inactivation protocol, as previously described (Eliot et al., 2017; McKinley et al., 2019a), and imaged on an Olympus X81 inverted microscope (20X magnification) with a motorized stage. For histological analysis, slides were processed and stained for hematoxylin and eosin and beta-catenin using standard approaches. Blind scoring was conducted by a pathologist (Dr. Kay Washington) using brightfield microscopy and a standard grading scale for dysplasia.

Murine Organoid Formation Assay

Organoids derived from Lrig1 and Mist1 tumors were dissociated using TrypLE Express. Cell pellets were resuspended in matrigel and seeded in 25 µL/well in a 24-well plate with 500 µL of Mouse Intesticult (STEMCELL Technologies) media. After one week, the number of organoids was counted using the GelCount™ system (Oxford Optronix). The number of organoids formed in each well was normalized to the number of single cells plated to determine organoid formation rate. Results were tabulated and plotted using Prism 9 (GraphPad) with unpaired t test.

Murine Organoid Antigen Processing and Presentation Assay

Organoids were formed and cultured for one week in Matrigel and Mouse Intesticult media. They were collected and reseeded without Matrigel in media with 100 µg/mL DQ-Ovalbumin (Thermo Fisher Scientific) for approximately 24 hours. After 24 hours, organoids were fixed, stained overnight with antibodies against GFP and Ia/Ie-AF647 (1:100; Biolegend), and analyzed using a BD LSRII 5-laser flow cytometer. Flow data were analyzed using Cytobank [234].

Murine T cell Activation Assay

Naïve OTII cells were isolated from the spleen of 8–10-week-old OT-II mice. Cells were purified using the naïve CD4+ T Cell Isolation Kit (STEMCELL Technologies) following manufacturer's protocol. CD11c+ DCs were isolated using MagniSort™ Mouse CD11c Positive Selection Kit (Thermo Fisher) per manufacturer's recommendations. Murine Organoids were dissociated with TrypLE containing 10 µM Y-27632 for 15 minutes at 37 °C while shaking. Cells were counted using Bio-RAD TC20 automated cell counter for use in the antigen presentation assay.

To track T cell proliferation, naïve CD4+ OTII T cells were labeled using 5 mM CellTrace Violet (Thermo Fisher) by incubating for 20 minutes at 37°C, 5% CO2 in PBS and then an equal volume of T cell media containing serum was added and incubated an additional 5 minutes at 37°C, 5% CO2 to quench free dye. $5x10^4$ labeled OTII CD4+ T cells were plated in a 96-well round bottom plate with $2.5x10^5$ organoid-dissociated single cells (without Matrigel) or 2.5x105 CD11c+ DCs and in the presence or absence of 50 µg/mL ovalbumin peptide (Anaspec), spun at 350 x g for 5 minutes and then incubated at 37°C, 5% CO2 for 72 hours. Following co-culture, cells were washed with PBS, stained with an antibody cocktail and assessed via flow cytometry. Wells containing cells were pipetted up and down to resuspend all cells and placed in 5mL Falcon™ Round-Bottom Polystyrene Tubes. These were centrifuged briefly at 350 x g for 3 minutes at 4°C, washed in FACS buffer (PBS w/o $Ca^{2+}Mg^{2+}$, 2% FBS, 2 mM EDTA), and resuspended in 100 µL

FACS buffer containing the antibody cocktail and stained for 15 min at 4°C. Cells were spun down as before, washed in FACS buffer, and were resuspended in 250 µL of FACS buffer and kept on ice until acquired on a 4-laser Fortessa. Cytometry data analysis was done using FlowJo v10 software and T cell proliferation results were tabulated and plotted in GraphPad Prism 9 using ANOVA with post-hoc Tukey tests. This protocol was adapted from [196].

MxIF, Single-cell Segmentation and Image Analysis

Cell segmentation was accomplished using the MANDO pipeline [235]. Briefly, random forest pixel classification on manually annotated images was used to define tissue and subcellular regions in each image. An initial watershed segmentation using cell nuclei as seed points and the learned cell membranes as boundaries was followed by re-segmentation of objects containing internal cell membranes. For every identified cell, image intensities for each marker were then calculated as well as morphological features such as cell size and location. For quantifying marker positive cells in MxIF, we fitted linear mixed effects models on the logit transformed cell proportions within epithelial or stromal tissue compartments. We estimated the proportion of marker positive cells within each compartment, by dividing the number of marker positive cells by the total number of cells within the tissue compartment. We added ½ to the numerator and denominator of the proportion to accommodate zero proportions; this is equivalent to a Bayesian estimator for the proportions using a noninformative beta prior. We fit the logit transformed proportions using a linear mixed effects model with an interaction between tissue compartment (epithelium/stromal), tissue type (AD/SSL), and a random effect for slide to account for the correlation between regions on a slide [236]. We estimated differences between tumor types within each tissue compartment using emmeans [237]. We computed false discovery rate (FDR) adjusted p-values using Benjamini-Hochberg. For murine tissue, tumor areas were established by a beta-Catenin mask and cell counts for image quantification were determined the same way as human tissues.

MxIF and MxIHC, Pixel-based Image Quantification

MATLAB was initially used to create masks to mark positive pixels of each cell type marker from MxIF images [238]. The tumor region was divided into an epithelial region (masked by beta-Catenin, pan-Cytokeratin, and NaKATPase expression) and a stromal region (tumor mask minus the epithelial mask). An overlay of OLFM4, MUC5AC, and PANCK was used as a guide for identifying stem (OLFM4+) and metaplastic (MUC5AC+) epithelial (PANCK+) regions, which were then manually demarcated. Each region was validated by quantifying MUC5AC and OLFM4 positive pixels within the regions. Cell types were defined by combinations of marker masks; for example, CD4+ T cells were defined by intersecting CD4 and CD3 pixel masks. On the other hand, CD8+ T cells were defined using the CD8 marker. We then calculated the fraction of pixels occupied by each cell type, normalized to the number of pixels of each tumor region. For example, a ratio of intraepithelial CD8+ cells to stromal CD4+ T cells was calculated from two sets of values calculated in this way. The measurements from all regions of the same type within each tumor was used to calculate a mean value; thus, each patient is a biological replicate. One-way ANOVA with Dunnett post-test was used for statistical testing. For IHC images, a similar process was used, with whole tumor regions demarcated by tissue morphology using hematoxylin nuclear counterstain. Antibody stains (3,3'-diaminobenzidine - DAB or 3-amino-9-ethylcarbazole - AEC) were spectrally unmixed such that individual marker masks can be generated and quantified as above.

Chapter V - Discussion and Future Directions

Recreated from:

Chen, B., Scurrah, C. R., Mckinley, E. T., Simmons, A. J., Ramirez-Solano, M. A., Zhu, X., Markham, N. O., Heiser, C. N., Vega, P. N., ... Coffey, R., Shrubsole, M., & Lau, K. S. (2021). Differential pre-malignant programs and microenvironment chart distinct paths to malignancy in human colorectal polyps. Cell. https://doi.org/10.1016/j.cell.2021.11.031

Discussion

We present a multi-omic analysis of the two major subclasses of pre-malignant polyps in the human colon: ADs and SERs that include SSLs and HPs. While the oncogenesis, progression, and stem cell origin of ADs and MSS CRCs are well-substantiated from human data and mouse models [110,111,132], our atlas provides novel insights into the less-studied SERs by comparing their biology to ADs. AD and MSS CRC cells exhibit higher degrees of stemness compared to normal colonic stem cells, consistent with what is known about WNT dysregulation in these tumors [239,240]. Importantly, we present evidence that SSLs and other associated SERs arise from gastric metaplasia based on a pyloric gland gene signature expressed in SSCs. In SERs, $MUC5AC^+$ cells were observed extending from the surface mucosa to the crypt base, consistent with the notion that these lesions originate at the luminal surface and extend downwards. This pattern contrasts markedly with ADs, which arise from the stem cell compartment at the crypt base. These findings provide support for both a top-down (in SSLs) and bottom-up (in ADs) model of colorectal tumorigenesis, and may help reconcile a long unresolved debate in the field [241,242].

By definition, metaplasia is a process in which differentiated cells transition into cell types that are non-native to the tissue. Metaplasia often arises in response to damage of the epithelium, which activates a regenerative program to direct the conversion to reparative mucous-secreting lineages

resembling those of pyloric glands [243]. Metaplastic programs have been observed in the stomach (spasmolytic polypeptide-expressing metaplasia, SPEM) [244,245], pancreas (acinar to ductal metaplasia, ADM) [246], and small intestine (ulcer-associated lineage, UACL) [247–249]. In SERs, we observed mis-expression of genes found in the gastric pylorus, reversion to a fetal gene program, and loss of regional identity with reduced *CDX2* expression. In the postnatal gut, loss of *Cdx2* expression in mice results in a rostral shift of tissue identity with expression of gastric markers [152]. It is important to distinguish metaplastic loss of regional identity from dedifferentiation of committed cells into less differentiated and even stem cell state [250–253], because the latter still retains the identity of the original organ. We propose a new paradigm in which damage in the proximal colon, possibly from microbiota, initiates a regenerative process resulting in loss of *CDX2* expression, gastric metaplasia, and reversion to a more embryonal state. Response to damage may activate survival/proliferative signaling pathways that eventually selects for activating *BRAF* mutations in metaplastic cells. Reversion to a fetal developmental identity is a feature of WNT-independent tumorigenesis found in recent mouse models [254], which can be triggered by MAPK activation either genetically by *Braf* activating mutations, response to epithelial damage, or stress triggered by mismatch repair deficiency [255,256]. Critically, *Braf* mutations in mouse models must be accompanied by a "second hit", such as perturbation of TGF-β signaling, for tumor induction [154,254,256]. This "second hit" may be provided by signals from the microenvironment.

Methylation of the *CDX2* locus has been frequently observed in serrated tumors, potentially leading to its downregulation, and loss of *Cdx2* can provide the "second hit" in a serrated tumorigenesis model [154]. Increased methylation has been found to be dependent on extrinsic factors such as aging [257], consistent with the preponderance of *BRAFV600E* mutations in MSI-H CRCs in older individuals [258]. Shown more recently, microbial dysbiosis can also be an environmental trigger for hypermethylation [221]. Antibiotic suppression of the microbiota reduces colonic tumorigenesis in a *Braf* mutant model [256], whereas in another study, enterotoxigenic

Bacteroides fragilis (ETBF) infection is a required trigger for tumorigenesis in the proximal mid-colon in a *Braf* mutant mouse model [221]. In the latter report, the earliest events of the ETBF response in epithelial cells prior to tumor formation occur at the colonic mucosal surface, where colonic epithelial cells and luminal contents interact. The importance of the microbiota to this type of tumorigenesis is underscored by the co-occurrence of polymicrobial biofilms in ~90% of right-sided CRCs, which are enriched for serrated tumors, versus ~12% biofilm-positive left-sided CRCs [259]. Considering the crypt-to-lumen vertical axis of the colonic mucosa, differentiated cells at the luminal surface are exposed to the microbiota, are more susceptible to damage, and utilize repair mechanisms reliant on cellular plasticity. Conversely, stem cells residing in the crypt base are more protected from luminal stressors [98]. We speculate that conventional adenomatous and serrated tumorigenesis originate from fundamentally different mechanisms: the former from DNA replication-induced mutations in continually proliferating stem cells and the latter from damage and repair at the colonic surface triggered and maintained by foreign stressors in the luminal environment. Distinct origins of neoplastic cells then select for different mutational pathways required for tumorigenesis.

Our data support distinct origins of serrated and conventional tumors based on histological, genetic, and transcriptomic evidence. Several of our findings have significant clinical relevance. SSLs can be challenging to identify as the diagnosis is based on the presence of a single "architecturally distorted serrated crypt" as defined by the recently revised WHO classification [260]. Presently, there are no accepted molecular markers to aid in the diagnosis. Based on our findings, MUC5AC staining, coupled with the absence of CDX2 staining, may confirm the diagnosis of lesions suspicious for SSLs. In addition, the cytotoxic immune response in SSLs is observed to precede hypermutation in human tumors, which is consistent with recent mouse modeling showing the same order of events [221]. Hypermutation is a characteristic of MSI-H CRCs, and the resulting high neoantigen load is thought to be the critical driver of the cytotoxic microenvironment.

What then drives the cytotoxic immune response without hypermutation? Our human data, supported by mouse modeling experiments, implicate the differentiation status of neoplastic cells. Specifically, tumors arising from differentiated cells are more adept at antigen presentation and setting up an adaptive immune environment similar to that of an anti-microbial response. This concept is consistent with normal mouse intestine where antigen presentation by intestinal stem cells positively correlates with the degree of differentiation [196]**.** Luminal microbial antigen passage has also been shown to be orchestrated by colonic goblet cells that potentially develops from *Mist1*+ progenitors described in this study [261]. Strikingly, driving tumorigenesis via WNT pathway activation in non-stem cells is sufficient to promote a more cytotoxic immune environment. How tumor cells with a differentiated phenotype acquire and maintain immuno-stimulating properties remains to be determined. In contrast, acquisition of stem cell characteristics by MSI-H CRCs contributes to spatial intratumoral heterogeneity: metaplastic compartments retain their association with cytotoxic immune cells, and stem cell compartments become associated with immunosuppressive cells and signals. Colon cancer stem-like cells have been shown to downregulate their antigen-presentation machinery [262,263]. The degree to which MSI-H CRCs acquire stem-like properties is variable; future studies will be needed to determine whether acquisition of “stemness” in these cancers impacts the likelihood of an immunotherapeutic response. The top-down spatial organization, differentiated and metaplastic transcriptional program, and cytotoxic immune environment associated with SSLs may open novel strategies for interception of cancer progression, including better informed interval guidelines for surveillance, chemoprevention, or pre- and pro-biotic therapies.

Future Directions

Next-generation computational methods

The work presented in this dissertation was only made possible by the application of approaches beyond their originating disciplines. Complexity science is interdisciplinary in nature, and it is

evident that complex systems, regardless of scale, share generalizable properties that may be better understood through their joint study. A concrete example is the development of my analytical HCI framework using methods originating in single-cell transcriptomics; some of these methods, such as inflection point optimization, originate in fields with further degrees of separation such as electrical engineering and signal processing. Inverting this pattern, our understanding of biological systems may even stand to benefit from the generalization of HCI complex systems analyses. This may be done by borrowing HCI process mining methods, used to understand behavioral workflows, and applying them to mining biological pathways **(Figure 30)**. All that is required for this type of analysis in information systems is categorical time series data, typically logging user interactions. Arguably, biological pathways are also sequential executions of functional units, but instead of user clicks and roles, biological system functions may be executed through the expression of genes and coordination of translated proteins. Single-cell transcriptomics aptly fulfills the requirements for this type of process-oriented analysis, given that genes represent quantifiable, categorical variables and RNA velocity establishes a relative timeline of gene transcript expression. Thus, the theoretical output of process mining transcriptomes would be a process tree designating relative timelines of probabilistic gene expression, analogous to signaling pathway diagrams.

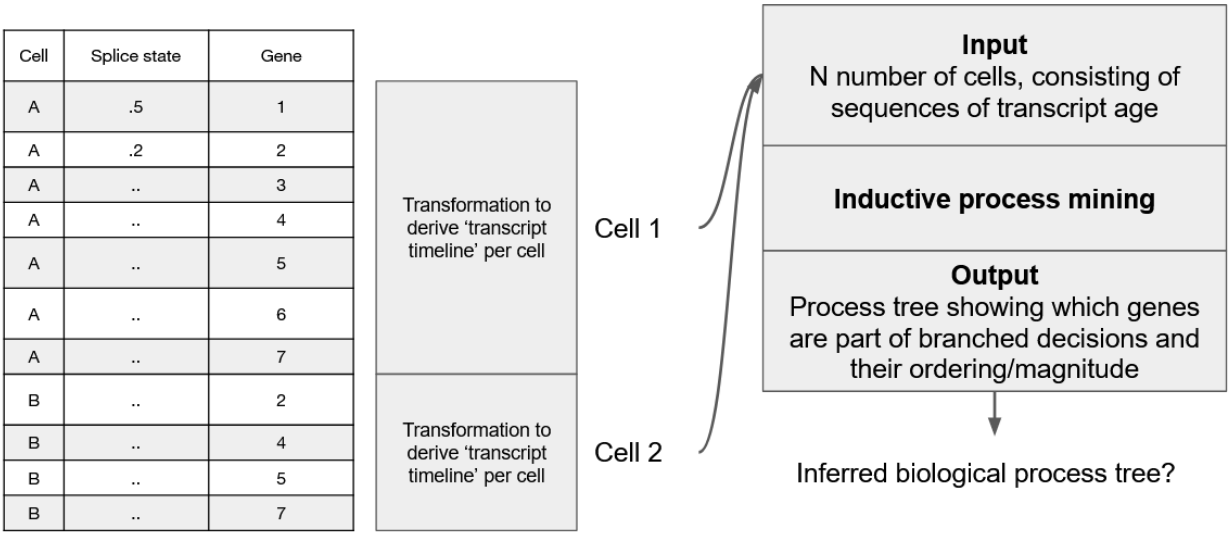


Figure 30. Hypothetical RNA velocity-based biological process mining.

In characterizing human tissue, one of our approaches introduced levels of abstraction by reframing gene expression into regulatory networks and describing the resulting clusters of cells. A strange loop arises through the application of the same clustering method to resulting regulatory networks themselves. Briefly, strange loops are a notable phenomenon which may arise in complex systems, where hierarchical levels of abstracted information become “tangled” in self-reference [264]. Illustrating this is the question of what is upstream and what is downstream of different layers of biological information: Where is the ground level in the central dogma of multicellular organisms, the DNA which originates translated RNA or the protein which enables the propagation of DNA? Here, an attempt to describe the gene regulation of single-cells, instead, becomes a top-level description of coordinated gene expression shared across all cells comprising the tissue. This unusual behavior of hierarchical abstraction remains poorly characterized in the context of biology and is likely tied with the emergent properties of complex systems. Importantly, this super-regulon concept is not unique among contemporary studies, with logically homologous frameworks published as Ecotypes and multicellular-immune hubs. Each of these frameworks involves the abstraction of regulatory programs across heterogeneous mixtures of single cell transcriptomes through the application of clustering to a feature space (genes and regulatory programs) as opposed to the observation space (cells and cell populations). Building on these ideas, next generation frameworks for exploring multicellular regulatory networks may involve yet another layer of abstraction.

Through the course of this thesis work, a major focus was the application of methods, whether they were existing or were novel developments. Given these applications, several methods should be highlighted for potential algorithmic improvement. Namely, CytoTRACE, p-Creode, and

SCENIC may likely be improved through the incorporation of regulon-based modalities, formal hierarchical models, and data-driven sampling methods. First, CytoTRACE is the method used in our characterization of stem and developmental potential which is trained on genes correlated with transcriptional diversity. Evidenced by our organoid experiments, this algorithm produces scores that are highly correlated with self-renewing properties; still, issues associated with the utilization of this algorithm originate in three aspects: its implementation in the R programming language, the confounding effects of batch integration, and memory utilization in large datasets. Given our established pipeline of transcriptomic quality control and scanpy-based analysis, the seamless integration of CytoTRACE would benefit from a python implementation with speedups drawing from the AnnData data structure and python parallelization libraries. Batch integration and memory utilization may be jointly addressed simultaneously through model training on regulons as opposed to genes. This is due to the batch-robust properties of the regulon feature extraction and the decreased feature load incorporated into the CytoTRACE regression model, yielding a lighter-weight calculation.

The benefit of these CytoTRACE improvements would propagate to our implementation of p-Creode as well, as the detection of root node endstates are typically associated with high CytoTRACE scores since stem-like cells are often the progenitors of a developmental process. Albeit a user input-heavy process, core functionality of p-Creode is its denoising process implemented through density-based down-sampling. With the recent publication of multiple intelligent sampling methods, namely Geosketch and Hopper, this denoising process may be automated in a streamline manner. Such methods, instead of relying on manually set density thresholds, learn the transcriptomic manifold and selectively sample from each of its regions, resulting in the balanced detection of rare and common cell types. Additionally, newer modalities of graphing libraries, along with their features, would improve the runtime and capabilities of p-Creode. Examples of this are the graph-tool library and the schist python packages, acting as

wrappers around C++ frameworks. These modalities allow for up to a 10-times, parallelized speedup of several graph-based calculations which may be used alongside other Python parallel processing implementations. By speeding up the runtime and sampling quality of each p-Creode run, more runs of higher quality may be incorporated into the ensemble of graphs for scoring. Outside of speedups, these graph libraries offer alternative graph architectures for trajectory inference, such as relative neighborhood graphs or stochastic block models.

Impact on understanding the human GI tract

Following the biological intuitions presented by my work published in Cell, the process of metaplasia in the human gastrointestinal tract may be more common than previously thought. The work described in this thesis includes a single-cell resolution atlas of human adenomas and serrated polyps, where serrated polyps arise from metaplasia as opposed to adenomatous stem cell expansion. Importantly, this work shows that cytotoxic immunity in serrated polyps occurs independently of hypermutation, and instead such distinct immune microenvironments track tumor cell- differentiation states. Our defined regulatory networks and gene signatures, in the form of regulon weighting matrices and gene sets, represent tangible abstractions of biological processes involved in metaplasia; these are generalizable for the characterization of any set of single-cell transcriptomes.

Extending the transcriptomic quality control pipeline we devised during the creation of this cell atlas, the deep characterization of biological processes occurring within dying cells may lead to novel methods of droplet classification. For example apoptosis, anoikis, and netosis are three tightly regulated processes in which a pathway of genes initiates the dissolution of cells. Once characterized in the context of single-cell transcriptomics, regulatory networks and gene sets may then be generalized for the detection of undesirable transcriptomes during the quality control process. Alongside this, the understanding of pervasively expressed mitochondrial or ribosomal

genes may lay the foundation for the subcategorization of previously defined cell states, like those involved in stem cell regeneration and proliferation.

By validating that T cell proliferation and the presentation of antigen processing machinery vary depending on tumor cell antigenic properties, our work refines the scope for future studies involving T cell regulation. Though macrophage and other myeloid-lineage cell type interactions were not the focus of our study, the investigation of some properties, such as macrophage localization and antigen processing, stand to benefit as our atlas represents an early snapshot during tumorigenesis. Inflammatory processes which may be regulated through macrophages, such as the IL-1R response, could be investigated for potential roles in crypt regeneration using our atlas as a starting point. In a similar vein, a shortcoming of this work is its lack of -omics level spatial analysis, which would benefit the understanding of juxtacrine signaling in the microenvironment. Stromal characterization through cell-cell interaction methods, especially those incorporating the gestalt of signaling cascades and their regulatory networks, still need to be defined alongside paired spatial transcriptomics.

Metaplasia, and closely related biological processes such as inflammation and stress or damage response, have been nominally characterized before in the context of the lower GI tract. Using multiple layers of molecular information, Nyström and co-authors described inter-cryptal goblet cells alongside noncanonical goblet cells which resembled the mucin-producing SSCs observed in colonic polyps. These similarities were evidenced by the expression of genes related to inflammation and damage response, as abstracted through gene set enrichment analysis. Further investigation may yield a shared process linking the behavior of these enterocyte-like goblet cells to a disruption of homeostasis that could result in serrated polyps or eventually MSI-H cancer.

References

1. Kitano, H. Systems biology: A brief overview. *Science (80-.).* **295**, 1662–1664 (2002).
2. Bertalanffy, L. Von. An outline of general system theory. *Br. J. Philos. Sci.* **1**, 134–165 (1950).
3. Bar-Yam, Y. General Features of Complex Systems. *Knowl. Manag. Organ. Intell. Learn. Complex.* **I**, 1–10 (1997).
4. Liu, Y. Y. & Barabási, A. L. Control principles of complex systems. *Rev. Mod. Phys.* **88**, 035006 (2016).
5. Gardner, M. Mathematical Games. *Sci. Am.* **220**, 116–120 (1969).
6. Wolfram, S. Cellular automata as models of complexity. *Nature* **311**, 419–424 (1984).
7. Wolfram, S. *A New Kind of Science*. (2002).
8. Aderem, A. Systems Biology: Its Practice and Challenges. *Cell* **121**, 511–513 (2005).
9. Lipsitz, L. A., Seniorlife, H. & Israel, B. Understanding Health Care as a Complex System: The Foundation for Unintended Consequences. *JAMA* **308**, 243–244 (2012).
10. Newman, M. E. J. Complex Systems: A Survey. *Am. J. Phys.* **79**, 800–810 (2011).
11. Barabási, A.-L. & Albert, R. Emergence of Scaling in Random Networks. *Science (80-.).* **286**, 509 LP – 512 (1999).
12. Bruggeman, F. J. & Westerhoff, H. V. The nature of systems biology. *Trends Microbiol.* **15**, 45–50 (2007).
13. Anderson, A. R. A. & Quaranta, V. Integrative mathematical oncology. *Nat. Rev. Cancer 2008 83* **8**, 227–234 (2008).
14. Scurrah, C. R., Simmons, A. J. & Lau, K. S. Single-Cell Mass Cytometry of Archived Human Epithelial Tissue for Decoding Cancer Signaling Pathways. *Methods Mol. Biol.* **1884**, 215–229 (2019).
15. Simmons, A. J. *et al.* Cytometry-based single-cell analysis of intact epithelial signaling reveals MAPK activation divergent from TNF-α-induced apoptosis in vivo. *Mol. Syst. Biol.*

11, 835 (2015).

16. Yamamura, S. *et al.* Single-cell microarray for analyzing cellular response. *Anal. Chem.* **77**, 8050–8056 (2005).
17. Kamme, F. *et al.* Single-Cell Microarray Analysis in Hippocampus CA1: Demonstration and Validation of Cellular Heterogeneity. *J. Neurosci.* **23**, 3607–3615 (2003).
18. Zhang, X. *et al.* Comparative Analysis of Droplet-Based Ultra-High-Throughput Single-Cell RNA-Seq Systems. *Mol. Cell* **73**, 130–142 (2019).
19. Klein, A. M. *et al.* Droplet Barcoding for Single-Cell Transcriptomics Applied. *Cell* **161**, 1187–1201 (2015).
20. Macosko, E. Z. *et al.* Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets. *Cell* **161**, 1202–1214 (2015).
21. Zheng, G. X. Y. *et al.* Massively parallel digital transcriptional profiling of single cells. *Nat. Commun.* **8**, 14049 (2017).
22. Southard-Smith, A. N. *et al.* Dual indexed library design enables compatibility of in-Drop single-cell RNA-sequencing with exAMP chemistry sequencing platforms. *BMC Genomics* (2020). doi:10.1186/s12864-020-06843-0
23. Islam, M., Chen, B., Spraggins, J. M., Kelly, R. T. & Lau, K. S. Use of Single-Cell -Omic Technologies to Study the Gastrointestinal Tract and Diseases, From Single Cell Identities to Patient Features. *Gastroenterology* (2020). doi:10.1053/j.gastro.2020.04.073
24. Cusanovich, D. A. *et al.* Multiplex single-cell profiling of chromatin accessibility by combinatorial cellular indexing. *Science (80-.).* (2015). doi:10.1126/science.aab1601
25. Buenrostro, J. D. *et al.* Single-cell chromatin accessibility reveals principles of regulatory variation. *Nature* (2015). doi:10.1038/nature14590
26. Lareau, C. A. *et al.* Droplet-based combinatorial indexing for massive-scale single-cell chromatin accessibility. *Nat. Biotechnol.* (2019). doi:10.1038/s41587-019-0147-6
27. Nagano, T. *et al.* Single-cell Hi-C reveals cell-to-cell variability in chromosome structure.

Nature (2013). doi:10.1038/nature12593

28. Ramani, V. *et al.* Sci-Hi-C: a single-cell Hi-C method for mapping 3D genome organization in large number of single cells. *bioRxiv* (2019). doi:10.1101/579573
29. Chen, X. *et al.* ATAC-see reveals the accessible genome by transposase-mediated imaging and sequencing. *Nat. Methods* (2016). doi:10.1038/nmeth.4031
30. Rotem, A. *et al.* Single-cell ChIP-seq reveals cell subpopulations defined by chromatin state. *Nat. Biotechnol.* (2015). doi:10.1038/nbt.3383
31. Kaya-Okur, H. S. *et al.* CUT&Tag for efficient epigenomic profiling of small samples and single cells. *Nat. Commun.* (2019). doi:10.1038/s41467-019-09982-5
32. Karemaker, I. D. & Vermeulen, M. Single-Cell DNA Methylation Profiling: Technologies and Biological Applications. *Trends in Biotechnology* (2018). doi:10.1016/j.tibtech.2018.04.002
33. Gravina, S., Dong, X., Yu, B. & Vijg, J. Single-cell genome-wide bisulfite sequencing uncovers extensive heterogeneity in the mouse liver methylome. *Genome Biol.* (2016). doi:10.1186/s13059-016-1011-3
34. Guo, H. *et al.* Profiling DNA methylome landscapes of mammalian cells with single-cell reduced-representation bisulfite sequencing. *Nat. Protoc.* (2015). doi:10.1038/nprot.2015.039
35. Han, L. *et al.* Bisulfite-independent analysis of CpG island methylation enables genome-scale stratification of single cells. *Nucleic Acids Res.* (2017). doi:10.1093/nar/gkx026
36. Mooijman, D., Dey, S. S., Boisset, J. C., Crosetto, N. & Van Oudenaarden, A. Single-cell 5hmC sequencing reveals chromosome-wide cell-to-cell variability and enables lineage reconstruction. *Nat. Biotechnol.* (2016). doi:10.1038/nbt.3598
37. Bormann, F. *et al.* Cell-of-Origin DNA Methylation Signatures Are Maintained during Colorectal Carcinogenesis. *Cell Rep.* (2018). doi:10.1016/j.celrep.2018.05.045
38. Litzenburger, U. M. *et al.* Single-cell epigenomic variability reveals functional cancer

heterogeneity. *Genome Biol.* (2017). doi:10.1186/s13059-016-1133-7

39. Chen, B. *et al.* Mining tasks and task characteristics from electronic health record audit logs with unsupervised machine learning. *J. Am. Med. Informatics Assoc.* (2021). doi:10.1093/jamia/ocaa338

40. Kroth, P. J. *et al.* Association of Electronic Health Record Design and Use Factors With Clinician Stress and Burnout. *JAMA Netw. open* (2019). doi:10.1001/jamanetworkopen.2019.9609

41. Adler-Milstein, J. & Huckman, R. S. The impact of electronic health record use on physician productivity. *Am. J. Manag. Care* (2013). doi:10.48009/4_iis_2020_1-8

42. Chen, L. *et al.* Racing Against the Clock: Internal Medicine Residents' Time Spent On Electronic Health Records. *J. Grad. Med. Educ.* **8**, 39 (2016).

43. Babbott, S. *et al.* Electronic medical records and physician stress in primary care: Results from the MEMO Study. *J. Am. Med. Informatics Assoc.* **21**, e100–e106 (2014).

44. Card, A. J. Physician Burnout: Resilience Training is Only Part of the Solution. *Ann. Fam. Med.* **16**, 267–270 (2018).

45. Robertson, S. L., Robinson, M. D. & Reid, A. Electronic Health Record Effects on Work-Life Balance and Burnout Within the I3 Population Collaborative. *J. Grad. Med. Educ.* **9**, 479 (2017).

46. Marc Overhage, J. & McCallie, D. Physician time spent using the electronic health record during outpatient encounters a descriptive study. *Ann. Intern. Med.* **172**, 169–174 (2020).

47. Rule, A., Chiang, M. F. & Hribar, M. R. Using electronic health record audit logs to study clinical activity: A systematic review of aims, measures, and methods. *Journal of the American Medical Informatics Association* (2020). doi:10.1093/jamia/ocz196

48. Cohen, G. R., Friedman, C. P., Ryan, A. M., Richardson, C. R. & Adler-Milstein, J. Variation in Physicians' Electronic Health Record Documentation and Potential Patient Harm from That Variation. *J. Gen. Intern. Med.* **34**, 2355–2367 (2019).

49. Monsen, C. B., Singh, A., Fellner, J. & Jackson, S. L. MEASURING PROVIDER EFFICIENCY IN EPIC: A PRELIMINARY MIXED-METHODS EXPLORATION OF THE" PROVIDER EFFICIENCY PROFILE". in *JOURNAL OF GENERAL INTERNAL MEDICINE* **33**, S253–S253 (SPRINGER 233 SPRING ST, NEW YORK, NY 10013 USA, 2018).

50. Arndt, B. G. *et al.* Tethered to the EHR: Primary Care Physician Workload Assessment Using EHR Event Log Data and Time-Motion Observations. *Ann. Fam. Med.* **15**, 419–426 (2017).

51. Sinsky, C. A. *et al.* Metrics for assessing physician activity using electronic health record log data. *J. Am. Med. Informatics Assoc.* **27**, 639–643 (2020).

52. Adler-Milstein, J., Adelman, J. S., Tai-Seale, M., Patel, V. L. & Dymek, C. EHR audit logs: A new goldmine for health services research? *J. Biomed. Inform.* **101**, 103343 (2020).

53. Chen, Y. *et al.* Inferring clinical workflow efficiency via electronic medical record utilization. in *AMIA annual symposium proceedings* **2015**, 416 (American Medical Informatics Association, 2015).

54. Chen, Y., Lorenzi, N., Nyemba, S., Schildcrout, J. S. & Malin, B. We work with them? Healthcare workers interpretation of organizational relations mined from electronic health records. *Int. J. Med. Inform.* **83**, 495–506 (2014).

55. Chen, Y., Patel, M. B., McNaughton, C. D. & Malin, B. A. Interaction patterns of trauma providers are associated with length of stay. *J. Am. Med. Informatics Assoc.* **25**, 790–799 (2018).

56. Chubb, J. R., Trcek, T., Shenoy, S. M. & Singer, R. H. Transcriptional Pulsing of a Developmental Gene. *Curr. Biol.* **16**, 1018–1025 (2006).

57. Bellman, R. E. Adaptive Control Processes. *Adapt. Control Process.* (1961). doi:10.1515/9781400874668/HTML

58. Welch, J. D., Hartemink, A. J. & Prins, J. F. SLICER: Inferring branched, nonlinear

cellular trajectories from single cell RNA-seq data. *Genome Biol.* **17**, 1–15 (2016).

59. Rizvi, A. H. *et al.* Single-cell topological RNA-seq analysis reveals insights into cellular differentiation and development. *Nat. Biotechnol.* **35**, 551–560 (2017).
60. Brennecke, P. *et al.* Accounting for technical noise in single-cell RNA-seq experiments. *Nat. Methods* **10**, 1093–1098 (2013).
61. Hotelling, H. Analysis of a complex of statistical variables into principal components. *J. Educ. Psychol.* **24**, 417–441 (1933).
62. Qiu, X. *et al.* Reversed graph embedding resolves complex single-cell trajectories. *Nat. Methods* **14**, 979–982 (2017).
63. Welch, J. D., Hartemink, A. J. & Prins, J. F. SLICER: inferring branched, nonlinear cellular trajectories from single cell RNA-seq data. *Genome Biol.* **17**, 106 (2016).
64. Furchtgott, L. A., Melton, S., Menon, V. & Ramanathan, S. Discovering sparse transcription factor codes for cell states and state transitions during development. *Elife* **6**, (2017).
65. Kim, J. & Marioni, J. C. Inferring the kinetics of stochastic gene expression from single-cell RNA-sequencing data. *Genome Biol.* **14**, R7 (2013).
66. Elowitz, M. B. Stochastic Gene Expression in a Single Cell. *Science (80-.).* **297**, 1183–1186 (2002).
67. Raj, A. & van Oudenaarden, A. Nature, nurture, or chance: stochastic gene expression and its consequences. *Cell* **135**, 216–26 (2008).
68. Herring, C. A. *et al.* Unsupervised Trajectory Analysis of Single-Cell RNA-Seq and Imaging Data Reveals Alternative Tuft Cell Origins in the Gut. *Cell Syst.* **6**, 37-51.e9 (2018).
69. Butler, A., Hoffman, P., Smibert, P., Papalexi, E. & Satija, R. Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nat. Biotechnol.* **36**, 411–420 (2018).

70. Taguchi, Y. Principal Component Analysis-Based Unsupervised Feature Extraction Applied to Single-Cell Gene Expression Analysis. in 816–826 (2018). doi:10.1007/978-3-319-95933-7_90

71. Chen, B., Herring, C. A. & Lau, K. S. pyNVR: investigating factors affecting feature selection from scRNA-seq data for lineage reconstruction. *Bioinformatics* (2019). doi:10.1093/bioinformatics/bty950

72. Rodriguez, A. & Laio, A. Clustering by fast search and find of density peaks. *Science (80-.).* **344**, 1492–1496 (2014).

73. van der Maaten, L. & Hinton, G. Visualizing Data using t-SNE. *J. Mach. Learn. Res.* **9**, 2579–2605 (2008).

74. Dobin, A. *et al.* STAR: Ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).

75. Liu, Q. *et al.* scRNABatchQC: multi-samples quality control for single cell RNA-seq data. *Bioinformatics* **35**, 5306–5308 (2019).

76. Petukhov, V. *et al.* dropEst: Pipeline for accurate estimation of molecular counts in droplet-based single-cell RNA-seq experiments. *Genome Biol.* **19**, 78 (2018).

77. Heiser, C. N. & Lau, K. S. A Quantitative Framework for Evaluating Single-Cell Data Structure Preservation by Dimensionality Reduction Techniques. *Cell Rep.* (2020). doi:10.1016/j.celrep.2020.107576

78. Saelens, W., Cannoodt, R., Todorov, H. & Saeys, Y. A comparison of single-cell trajectory inference methods. *Nat. Biotechnol.* (2019). doi:10.1038/s41587-019-0071-9

79. Herring, C. A., Chen, B., McKinley, E. T. & Lau, K. S. Single-Cell Computational Strategies for Lineage Reconstruction in Tissue Systems. *CMGH* (2018). doi:10.1016/j.jcmgh.2018.01.023

80. Chen, H. *et al.* Single-cell trajectories reconstruction, exploration and mapping of omics data with STREAM. *Nat. Commun.* **10**, 1903 (2019).

81. Zhou, Z., Xu, B., Minn, A. & Zhang, N. R. DENDRO: genetic heterogeneity profiling and subclone detection by single-cell RNA sequencing. *Genome Biol.* **21**, 10 (2020).

82. Lu, T. *et al.* Overcoming Expressional Drop-outs in Lineage Reconstruction from Single-Cell RNA-Sequencing Data. *Cell Rep.* **34**, 108589 (2021).

83. La Manno, G. *et al.* RNA velocity of single cells. *Nature* **560**, 494–498 (2018).

84. Bergen, V., Lange, M., Peidli, S., Wolf, F. A. & Theis, F. J. Generalizing RNA velocity to transient cell states through dynamical modeling. *Nat. Biotechnol. 2020 3812* **38**, 1408–1414 (2020).

85. Satija, R., Farrell, J. A., Gennert, D., Schier, A. F. & Regev, A. Spatial reconstruction of single-cell gene expression data. *Nat. Biotechnol.* (2015). doi:10.1038/nbt.3192

86. Kotliar, D. *et al.* Identifying gene expression programs of cell-type identity and cellular activity with single-cell RNA-Seq. *Elife* **8**, e43803 (2019).

87. Aibar, S. *et al.* SCENIC: Single-cell regulatory network inference and clustering. *Nat. Methods* (2017). doi:10.1038/nmeth.4463

88. Armingol, E., Officer, A., Harismendy, O. & Lewis, N. E. Deciphering cell–cell interactions and communication from gene expression. *Nat. Rev. Genet. 2020 222* **22**, 71–88 (2020).

89. Almet, A. A., Cang, Z., Jin, S. & Nie, Q. The landscape of cell–cell communication through single-cell transcriptomics. *Curr. Opin. Syst. Biol.* **26**, 12–23 (2021).

90. Efremova, M., Vento-Tormo, M., Teichmann, S. A. & Vento-Tormo, R. CellPhoneDB: inferring cell–cell communication from combined expression of multi-subunit ligand–receptor complexes. *Nat. Protoc. 2020 154* **15**, 1484–1506 (2020).

91. Wang, Y. *et al.* iTALK: an R Package to Characterize and Illustrate Intercellular Communication. *bioRxiv* 507871 (2019). doi:10.1101/507871

92. Luca, B. A. *et al.* Atlas of clinically distinct cell states and ecosystems across human solid tumors. *Cell* **184**, 5482-5496.e28 (2021).

93. Pelka, K. *et al.* Spatially organized multicellular immune hubs in human colorectal cancer.

Cell (2021). doi:10.1016/j.cell.2021.08.003

94. Browaeys, R., Saelens, W. & Saeys, Y. NicheNet: modeling intercellular communication by linking ligands to target genes. *Nat. Methods 2019 172* **17**, 159–162 (2019).

95. Azzouz, L. L. & Sharma, S. Physiology, Large Intestine. *StatPearls* (2021).

96. Lodish, H. *et al. Molecular Cell Biology*. (2013).

97. Müller, M. *et al.* Distal colonic transit is linked to gut microbiota diversity and microbial fermentation in humans with slow colonic transit. *Am. J. Physiol. - Gastrointest. Liver Physiol.* **318**, G361–G369 (2020).

98. Kaiko, G. E. *et al.* The Colonic Crypt Protects Stem Cells from Microbiota-Derived Metabolites. *Cell* **165**, 1708–1720 (2016).

99. Cheng, H. & Leblond, C. P. Origin, differentiation and renewal of the four main epithelial cell types in the mouse small intestine V. Unitarian theory of the origin of the four epithelial cell types. *Am. J. Anat.* **141**, 537–561 (1974).

100. Stappenbeck, T. S., Wong, M. H., Saam, J. R., Mysorekar, I. U. & Gordon, J. I. Notes from some crypt watchers: regulation of renewal in the mouse intestinal epithelium. *Curr. Opin. Cell Biol.* **10**, 702–709 (1998).

101. Barker, N., De Wetering, M. Van & Clevers, H. The intestinal stem cell. *Genes Dev.* **22**, 1856–1864 (2008).

102. Marshman, E., Booth, C. & Potten, C. S. The intestinal epithelial stem cell. *BioEssays* **24**, 91–98 (2002).

103. Kosinski, C. *et al.* Gene expression patterns of human colon tops and basal crypts and BMP antagonists as intestinal stem cell niche factors. *Proc. Natl. Acad. Sci.* **104**, 15418–15423 (2007).

104. Vinson, K. E., George, D. C., Fender, A. W., Bertrand, F. E. & Sigounas, G. The Notch pathway in colorectal cancer. *Int. J. Cancer* **138**, 1835–1842 (2016).

105. Du, H., Nie, Q. & Holmes, W. R. The Interplay between Wnt Mediated Expansion and

Negative Regulation of Growth Promotes Robust Intestinal Crypt Structure and Homeostasis. *PLOS Comput. Biol.* **11**, e1004285 (2015).

106. Clevers, H. & Batlle, E. EphB/EphrinB receptors and Wnt signaling in colorectal cancer. *Cancer Research* (2006). doi:10.1158/0008-5472.CAN-05-3849

107. Cadigan, K. M. & Waterman, M. L. TCF/LEFs and Wnt Signaling in the Nucleus. *Cold Spring Harb. Perspect. Biol.* **4**, a007906 (2012).

108. Du, L. *et al.* CD44 is of Functional Importance for Colorectal Cancer Stem Cells. *Clin. Cancer Res.* **14**, 6751–6760 (2008).

109. van der Flier, L. G., Haegebarth, A., Stange, D. E., van de Wetering, M. & Clevers, H. OLFM4 Is a Robust Marker for Stem Cells in Human Intestine and Marks a Subset of Colorectal Cancer Cells. *Gastroenterology* **137**, 15–17 (2009).

110. Barker, N. *et al.* Identification of stem cells in small intestine and colon by marker gene Lgr5. *Nature* **449**, 1003–1007 (2007).

111. Powell, A. E. *et al.* The Pan-ErbB Negative Regulator Lrig1 Is an Intestinal Stem Cell Marker that Functions as a Tumor Suppressor. *Cell* **149**, 146–158 (2012).

112. Altmann, G. G. Morphological observations on mucus-secreting nongoblet cells in the deep crypts of the rat ascending colon. *Am. J. Anat.* **167**, 95–117 (1983).

113. Sasaki, N. *et al.* Reg4+ deep crypt secretory cells function as epithelial niche for Lgr5+ stem cells in colon. *Proc. Natl. Acad. Sci. U. S. A.* **113**, E5399–E5407 (2016).

114. Poulin, E. J. *et al.* Using a new Lrig1 reporter mouse to assess differences between two Lrig1 antibodies in the intestine. *Stem Cell Res.* **13**, 422–430 (2014).

115. Barker, N., Van Oudenaarden, A. & Clevers, H. Identifying the Stem Cell of the Intestinal Crypt: Strategies and Pitfalls. *Cell Stem Cell* **11**, 452–460 (2012).

116. Elmentaite, R. *et al.* Cells of the human intestinal tract mapped across space and time. *Nat. 2021 5977875* **597**, 250–255 (2021).

117. Bigaeva, E., Uniken Venema, W. T. C., Weersma, R. K. & Festen, E. A. M.

Understanding human gut diseases at single-cell resolution. *Hum. Mol. Genet.* **29**, R51–R58 (2020).

118. Parikh, K. *et al.* Colonic epithelial cell diversity in health and inflammatory bowel disease. *Nat. 2019 5677746* **567**, 49–55 (2019).

119. De Santa Barbara, P., Van Den Brink, G. R. & Roberts, D. J. Development and differentiation of the intestinal epithelium. *Cell. Mol. Life Sci. C. 2003 607* **60**, 1322–1332 (2003).

120. Gerbe, F., Brulin, B., Makrini, L., Legraverend, C. & Jay, P. DCAMKL-1 Expression Identifies Tuft Cells Rather Than Stem Cells in the Adult Mouse Intestinal Epithelium. *Gastroenterology* **137**, 2179–2180 (2009).

121. McKinley, E. T. *et al.* Optimized multiplex immunofluorescence single-cell analysis reveals tuft cell heterogeneity. *JCI insight* (2017). doi:10.1172/jci.insight.93487

122. Gerbe, F. *et al.* Intestinal epithelial tuft cells initiate type 2 mucosal immunity to helminth parasites. *Nat. 2016 5297585* **529**, 226–230 (2016).

123. Howitt, M. R. *et al.* Tuft cells, taste-chemosensory cells, orchestrate parasite type 2 immunity in the gut. *Science (80-.).* **351**, 1329–1333 (2016).

124. Llosa, N. J. *et al.* The Vigorous Immune Microenvironment of Microsatellite Instable Colon Cancer Is Balanced by Multiple Counter-Inhibitory Checkpoints. *Cancer Discov.* **5**, 43–51 (2015).

125. Charles A Janeway, J., Travers, P., Walport, M. & Shlomchik, M. J. Immunobiology. *Immunobiology* 1–10 (2001).

126. Siegel, R. L., Miller, K. D. & Jemal, A. Cancer statistics, 2019. *CA. Cancer J. Clin.* (2019). doi:10.3322/caac.21551

127. Dienstmann, R. *et al.* Consensus molecular subtypes and the evolution of precision medicine in colorectal cancer. *Nat. Rev. Cancer* **17**, 79–92 (2017).

128. Guinney, J. *et al.* The consensus molecular subtypes of colorectal cancer. *Nat. Med.*

(2015). doi:10.1038/nm.3967

129. Ogino, S. & Goel, A. Molecular classification and correlates in colorectal cancer. *J. Mol. Diagnostics* **10**, 13–27 (2008).

130. Le, D. T. *et al.* PD-1 Blockade in Tumors with Mismatch-Repair Deficiency. *N. Engl. J. Med.* **372**, 2509–2520 (2015).

131. Mlecnik, B. *et al.* Integrative Analyses of Colorectal Cancer Show Immunoscore Is a Stronger Predictor of Patient Survival Than Microsatellite Instability. *Immunity* (2016). doi:10.1016/j.immuni.2016.02.025

132. Fearon, E. R. & Vogelstein, B. A genetic model for colorectal tumorigenesis. *Cell* (1990). doi:10.1016/0092-8674(90)90186-I

133. Crockett, S. D. & Nagtegaal, I. D. Terminology, Molecular Features, Epidemiology, and Management of Serrated Colorectal Neoplasia. *Gastroenterology* **157**, 949-966.e4 (2019).

134. Thorstensen, L. *et al.* Genetic and epigenetic changes of components affecting the WNT pathway in colorectal carcinomas stratified by microsatellite instability. *Neoplasia* **7**, 99–108 (2005).

135. Leggett, B. & Whitehall, V. Role of the Serrated Pathway in Colorectal Cancer Pathogenesis. *Gastroenterology* **138**, 2088–2100 (2010).

136. Yang, S., Farraye, F. A., Mack, C., Posnik, O. & O'Brien, M. J. BRAF and KRAS Mutations in Hyperplastic Polyps and Serrated Adenomas of the Colorectum: Relationship to Histology and CpG Island Methylation Status. *Am. J. Surg. Pathol.* **28**, (2004).

137. Jones, S. *et al.* Comparative lesion sequencing provides insights into tumor evolution. *Proc. Natl. Acad. Sci.* **105**, 4283 LP – 4288 (2008).

138. Markowitz, S. D. & Bertagnolli, M. M. Molecular origins of cancer: Molecular basis of colorectal cancer. *N. Engl. J. Med.* **361**, 2449–2460 (2009).

139. Heiser, C. N., Wang, V. M., Chen, B., Hughey, J. J. & Lau, K. S. Automated quality control and cell identification of droplet-based single-cell data using dropkick. *Genome Res.* gr.271908.120 (2021). doi:10.1101/GR.271908.120

140. Van de Sande, B. *et al.* A scalable SCENIC workflow for single-cell gene regulatory network analysis. *Nat. Protoc.* **15**, 2247–2276 (2020).

141. Imajo, M., Ebisuya, M. & Nishida, E. Dual role of YAP and TAZ in renewal of the intestinal epithelium. *Nat. Cell Biol. 2014 171* **17**, 7–19 (2014).

142. Lili, L. N. *et al.* Claudin-based barrier differentiation in the colonic epithelial crypt niche involves Hopx/Klf4 and Tcf7l2/Hnf4-α cascades. *http://dx.doi.org/10.1080/21688370.2016.1214038* **4**, (2016).

143. Paquet-Fifield, S. *et al.* Tight Junction Protein Claudin-2 Promotes Self-Renewal of Human Colorectal Cancer Stem-like Cells. *Cancer Res.* **78**, 2925–2938 (2018).

144. Gulati, G. S. *et al.* Single-cell transcriptional diversity is a hallmark of developmental potential. *Science (80-.).* (2020). doi:10.1126/science.aax0249

145. Korinek, V. *et al.* Constitutive transcriptional activation by a β-catenin-Tcf complex in APC(-/-) colon carcinoma. *Science (80-.).* (1997). doi:10.1126/science.275.5307.1784

146. Heilig, R. *et al.* The Gasdermin-D pore acts as a conduit for IL-1β secretion in mice. *Eur. J. Immunol.* **48**, 584–592 (2018).

147. Kim, J. H. *et al.* Gastric-type expression signature in serrated pathway–associated colorectal tumors. *Hum. Pathol.* **46**, 643–656 (2015).

148. Lee, S. H. *et al.* Up-regulation of Aquaporin 5 Defines Spasmolytic Polypeptide-Expressing Metaplasia and Progression to Incomplete Intestinal Metaplasia. *Cell. Mol. Gastroenterol. Hepatol.* **13**, 199–217 (2021).

149. Min, J. *et al.* Heterogeneity and dynamics of active Kras-induced dysplastic lineages from mouse corpus stomach. *Nat. Commun.* **10**, 5549 (2019).

150. Quante, M. *et al.* Bile Acid and Inflammation Activate Gastric Cardia Stem Cells in a

Mouse Model of Barrett-Like Metaplasia. *Cancer Cell* **21**, 36–51 (2012).

151. Tsai, J.-H. *et al.* Aberrant expression of annexin A10 is closely related to gastric phenotype in serrated pathway to colorectal carcinoma. *Mod. Pathol. 2015 282* **28**, 268–278 (2014).

152. Balbinot, C. *et al.* The Cdx2 homeobox gene suppresses intestinal tumorigenesis through non–cell-autonomous mechanisms. *J. Exp. Med.* **215**, 911–926 (2018).

153. Stringer, E. J. *et al.* Cdx2 determines the fate of postnatal intestinal endoderm. *Development* **139**, 465–474 (2012).

154. Tong, K. *et al.* Degree of Tissue Differentiation Dictates Susceptibility to BRAF-Driven Colorectal Cancer. *Cell Rep.* **21**, 3833–3845 (2017).

155. Park, Y. K. *et al.* Gene expression profile analysis of mouse colon embryonic development. *Genesis* (2005). doi:10.1002/gene.20088

156. Bjeldanes, L. F., Kim, J. Y., Grose, K. R., Bartholomew, J. C. & Bradfield, C. A. Aromatic hydrocarbon responsiveness-receptor agonists generated from indole-3-carbinol in vitro and in vivo: comparisons with 2,3,7,8-tetrachlorodibenzo-p-dioxin. *Proc. Natl. Acad. Sci.* **88**, 9543–9547 (1991).

157. Grizotte-Lake, M. *et al.* Commensals Suppress Intestinal Epithelial Cell Retinoic Acid Synthesis to Regulate Interleukin-22 Activity and Prevent Microbial Dysbiosis. *Immunity* **49**, 1103-1115.e6 (2018).

158. Iyer, N. & Vaishnava, S. Vitamin A at the interface of host–commensal–pathogen interactions. *PLOS Pathog.* **15**, e1007750 (2019).

159. Zhang, L., Nichols, R. G. & Patterson, A. D. The aryl hydrocarbon receptor as a moderator of host-microbiota communication. *Curr. Opin. Toxicol.* **2**, 30–35 (2017).

160. Lukonin, I. *et al.* Phenotypic landscape of intestinal organoid regeneration. *Nature* **586**, 275–280 (2020).

161. Paweletz, C. P. *et al.* Loss of Annexin 1 Correlates with Early Onset of Tumorigenesis in

Esophageal and Prostate Carcinoma. *Cancer Res.* **60**, 6293 LP – 6297 (2000).

162. Yui, S. *et al.* YAP/TAZ-Dependent Reprogramming of Colonic Epithelium Links ECM Remodeling to Tissue Regeneration. *Cell Stem Cell* **22**, 35-49.e7 (2018).
163. Ayyaz, A. *et al.* Single-cell transcriptomes of the regenerating intestine reveal a revival stem cell. *Nature* **569**, 121–125 (2019).
164. Murata, K. *et al.* Ascl2-Dependent Cell Dedifferentiation Drives Regeneration of Ablated Intestinal Stem Cells. *Cell Stem Cell* **26**, 377-390.e6 (2020).
165. Raudvere, U. *et al.* g:Profiler: a web server for functional enrichment analysis and conversions of gene lists (2019 update). *Nucleic Acids Res.* **47**, W191–W198 (2019).
166. Antushevich, H. Interplays between inflammasomes and viruses, bacteria (pathogenic and probiotic), yeasts and parasites. *Immunol. Lett.* **228**, 1–14 (2020).
167. Elinav, E. *et al.* NLRP6 Inflammasome Regulates Colonic Microbial Ecology and Risk for Colitis. *Cell* **145**, 745–757 (2011).
168. Levy, M. *et al.* Microbiota-Modulated Metabolites Shape the Intestinal Microenvironment by Regulating NLRP6 Inflammasome Signaling. *Cell* **163**, 1428–1443 (2015).
169. Man, S. M. Inflammasomes in the gastrointestinal tract: infection, cancer and gut microbiota homeostasis. *Nat. Rev. Gastroenterol. Hepatol.* **15**, 721–737 (2018).
170. Ansari, I. *et al.* The microbiota programs DNA methylation to control intestinal homeostasis and inflammation. *Nat. Microbiol.* (2020). doi:10.1038/s41564-019-0659-3
171. Singh, P., Rawat, A., Alwakeel, M., Sharif, E. & Al Khodor, S. The potential role of vitamin D supplementation as a gut microbiota modifier in healthy individuals. *Sci. Rep.* **10**, 21641 (2020).
172. Thomas, R. L. *et al.* Vitamin D metabolites and the gut microbiome in older men. *Nat. Commun.* **11**, 5997 (2020).
173. Rex, D. K. *et al.* Serrated Lesions of the Colorectum: Review and Recommendations From an Expert Panel. *Off. J. Am. Coll. Gastroenterol.* | *ACG* **107**, (2012).

174. Davenport, J. R. *et al.* Modifiable lifestyle factors associated with risk of sessile serrated polyps, conventional adenomas and hyperplastic polyps. *Gut* **67**, 456 LP – 465 (2018).

175. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* (2009). doi:10.1093/bioinformatics/btp324

176. Tarasov, A., Vilella, A. J., Cuppen, E., Nijman, I. J. & Prins, P. Sambamba: Fast processing of NGS alignment formats. *Bioinformatics* (2015). doi:10.1093/bioinformatics/btv098

177. Broad Institute. Picard toolkit. *Broad Institute, GitHub repository* (2019).

178. der Auwera, G. A. *Genomics in the cloud : using Docker, GATK, and WDL in Terra*. *Genomics in the cloud : using Docker, GATK, and WDL in Terra* (2020).

179. Van der Auwera, G. A. *et al.* From fastQ data to high-confidence variant calls: The genome analysis toolkit best practices pipeline. *Curr. Protoc. Bioinforma.* (2013). doi:10.1002/0471250953.bi1110s43

180. Banerjee, A. *et al.* Succinate Produced by Intestinal Microbes Promotes Specification of Tuft Cells to Suppress Ileal Inflammation. *Gastroenterology* **159**, 2101-2115.e5 (2020).

181. Chen, B., Ramirez-Solano, M. A., Heiser, C. N., Liu, Q. & Lau, K. S. Processing single-cell RNA-seq data for dimension reduction-based analyses using open-source tools. *STAR Protoc.* (2021). doi:10.1016/j.xpro.2021.100450

182. Wolf, F. A., Angerer, P. & Theis, F. J. SCANPY: large-scale single-cell gene expression data analysis. *Genome Biol.* **19**, 15 (2018).

183. McInnes, L., Healy, J., Saul, N. & Großberger, L. UMAP: Uniform Manifold Approximation and Projection. *J. Open Source Softw.* (2018). doi:10.21105/joss.00861

184. Poplin, R. *et al.* Scaling accurate genetic variant discovery to tens of thousands of samples. *bioRxiv* (2017). doi:10.1101/201178

185. Wang, K., Li, M. & Hakonarson, H. ANNOVAR: Functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res.* (2010).

doi:10.1093/nar/gkq603

186. Yang, H. & Wang, K. Genomic variant annotation and prioritization with ANNOVAR and wANNOVAR. *Nat. Protoc.* (2015). doi:10.1038/nprot.2015.105

187. Rocklin, M. Dask: Parallel Computation with Blocked algorithms and Task Scheduling. in *Proceedings of the 14th Python in Science Conference* (2015). doi:10.25080/majora-7b98e3ed-013

188. Traag, V. A., Waltman, L. & van Eck, N. J. From Louvain to Leiden: guaranteeing well-connected communities. *Sci. Rep.* **9**, 5233 (2019).

189. Shannon, P. *et al.* Cytoscape: A software Environment for integrated models of biomolecular interaction networks. *Genome Res.* (2003). doi:10.1101/gr.1239303

190. Harris, C. R. *et al.* Array programming with NumPy. *Nature* **585**, 357–362 (2020).

191. Korsunsky, I. *et al.* Fast, sensitive and accurate integration of single-cell data with Harmony. *Nat. Methods* (2019). doi:10.1038/s41592-019-0619-0

192. Caswell, T. A. *et al.* matplotlib/matplotlib: REL: v3.1.1. (2019). doi:10.5281/ZENODO.3264781

193. Waskom, M. *et al.* mwaskom/seaborn: v0.11.0 (Sepetmber 2020). (2020). doi:10.5281/ZENODO.4019146

194. Terpilowski, M. scikit-posthocs: Pairwise multiple comparison tests in Python. *J. Open Source Softw.* (2019). doi:10.21105/joss.01169

195. Varoquaux, G. *et al.* Scikit-learn. *GetMobile Mob. Comput. Commun.* (2017). doi:10.1145/2786984.2786995

196. Biton, M. *et al.* T Helper Cell Cytokines Modulate Intestinal Stem Cell Renewal and Differentiation. *Cell* **175**, 1307-1320.e22 (2018).

197. Mootha, V. K. *et al.* PGC-1α-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nat. Genet.* (2003). doi:10.1038/ng1180

198. Subramanian, A. *et al.* Gene set enrichment analysis: A knowledge-based approach for

interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. U. S. A.* (2005). doi:10.1073/pnas.0506580102

199. Reback, J. *et al.* pandas-dev/pandas: Pandas 1.0.3. (2020). doi:10.5281/zenodo.3715232
200. Eide, P. W., Bruun, J., Lothe, R. A. & Sveen, A. CMScaller: an R package for consensus molecular subtyping of colorectal cancer pre-clinical models. *Sci. Rep.* **7**, 16618 (2017).
201. La Manno, G. *et al.* RNA velocity of single cells. *Nature* **560**, 494–498 (2018).
202. Hartigan, J. A. & Wong, M. A. A k-means clustering algorithm. *Appl. Stat.* (1979). doi:10.2307/2346830
203. Stuart, T. *et al.* Comprehensive integration of single cell data. *bioRxiv* (2018). doi:10.1101/460147
204. Stoler, N. & Nekrutenko, A. Sequencing error profiles of Illumina sequencing instruments. *NAR Genomics Bioinforma.* **3**, (2021).
205. Fox, E. J., Reid-Bayliss, K. S., Emond, M. J. & Loeb, L. A. Accuracy of Next Generation Sequencing Platforms. *Next Gener. Seq. Appl.* **1**, (2014).
206. Skelly, D. A., Johansson, M., Madeoy, J., Wakefield, J. & Akey, J. M. A powerful and flexible statistical framework for testing hypotheses of allele-specific gene expression from RNA-seq data. *Genome Res.* **21**, 1728–1737 (2011).
207. Xiong, L. *et al.* SCALE method for single-cell ATAC-seq analysis via latent feature extraction. *Nat. Commun.* **10**, 4576 (2019).
208. Lee, H. O. *et al.* Lineage-dependent gene expression programs influence the immune landscape of colorectal cancer. *Nat. Genet.* (2020). doi:10.1038/s41588-020-0636-z
209. Chi, X.-Z. *et al.* Runt-related transcription factor RUNX3 is a target of MDM2-mediated ubiquitination. *Cancer Res.* **69**, 8111–8119 (2009).
210. Feng, X. *et al.* Transcription factor Foxp1 exerts essential cell-intrinsic regulation of the quiescence of naive T cells. *Nat. Immunol.* **12**, 544–550 (2011).

211. Liao, G.-B. *et al.* Regulation of the master regulator FOXM1 in cancer. *Cell Commun. Signal.* **16**, 57 (2018).

212. Chang, K. *et al.* Colorectal premalignancy is associated with consensus molecular subtypes 1 and 2. *Ann. Oncol.* **29**, 2061–2067 (2018).

213. Komor, M. A. *et al.* Consensus molecular subtype classification of colorectal adenomas. *J. Pathol.* **246**, 266–276 (2018).

214. Muzny, D. M. *et al.* Comprehensive molecular characterization of human colon and rectal cancer. *Nature* **487**, 330–337 (2012).

215. Le, D. T. *et al.* Mismatch repair deficiency predicts response of solid tumors to PD-1 blockade. *Science* **357**, 409–413 (2017).

216. Moon, C., VanDussen, K. L., Miyoshi, H. & Stappenbeck, T. S. Development of a primary mouse intestinal epithelial cell monolayer culture system to evaluate factors that modulate IgA transcytosis. *Mucosal Immunol.* **7**, 818–828 (2014).

217. Pabst, O. & Slack, E. IgA and the intestinal microbiota: the importance of being specific. *Mucosal Immunol.* **13**, 12–21 (2020).

218. Hieshima, K. *et al.* Molecular cloning of a novel human CC chemokine liver and activation- regulated chemokine (LARC) expressed in liver. Chemotactic activity for lymphocytes and gene localization on chromosome 2. *J. Biol. Chem.* (1997). doi:10.1074/jbc.272.9.5846

219. Nelson, R. T. *et al.* Genomic Organization of the CC Chemokine MIP-3α/CCL20/LARC/EXODUS/SCYA20, Showing Gene Structure, Splice Variants, and Chromosome Localization. *Genomics* **73**, 28–37 (2001).

220. Baruch, E. N. *et al.* Fecal microbiota transplant promotes response in immunotherapy-refractory melanoma patients. *Science (80-.).* **371**, 602–609 (2021).

221. Shields, C. E. D. *et al.* Bacterial-Driven Inflammation and Mutant BRAF Expression Combine to Promote Murine Colon Tumorigenesis That Is Sensitive to Immune

Checkpoint Therapy. *Cancer Discov.* **11**, 1792–1807 (2021).

222. Bonneville, M. *et al.* Intestinal intraepithelial lymphocytes are a distinct set of gamma delta T cells. *Nature* **336**, 479–481 (1988).

223. Schön, M. P. *et al.* Mucosal T Lymphocyte Numbers Are Selectively Reduced in Integrin α_E (CD103)-Deficient Mice. *J. Immunol.* **162**, 6641 LP – 6649 (1999).

224. Carlin, L. M. *et al.* Secretion of IFN-γ and not IL-2 by anergic human T cells correlates with assembly of an immature immune synapse. *Blood* **106**, 3874–3879 (2005).

225. TF, G., D, Q., P, F. & FW, F. Anergic T-lymphocyte clones have altered inositol phosphate, calcium, and tyrosine kinase signaling pathways. *Proc. Natl. Acad. Sci. U. S. A.* **91**, 38–42 (1994).

226. Wells, A. D., Walsh, M. C., Sankaran, D. & Turka, L. A. T Cell Effector Function and Anergy Avoidance Are Quantitatively Linked to Cell Division. *J. Immunol.* **165**, 2432–2443 (2000).

227. Fife, B. T. *et al.* Interactions between PD-1 and PD-L1 promote tolerance by blocking the TCR–induced stop signal. *Nat. Immunol. 2009 1011* **10**, 1185–1192 (2009).

228. Greenwald, R. J., Boussiotis, V. A., Lorsbach, R. B., Abbas, A. K. & Sharpe, A. H. CTLA-4 Regulates Induction of Anergy In Vivo. *Immunity* **14**, 145–155 (2001).

229. Duckworth, A., Glenn, M., Slupsky, J. R., Packham, G. & Kalakonda, N. Variable induction of PRDM1 and differentiation in chronic lymphocytic leukemia is associated with anergy. *Blood* **123**, 3277–3285 (2014).

230. Anderson, A. C., Joller, N. & Kuchroo, V. K. Lag-3, Tim-3, and TIGIT: Co-inhibitory Receptors with Specialized Functions in Immune Regulation. *Immunity* **44**, 989–1004 (2016).

231. Tsujikawa, T. *et al.* Quantitative Multiplex Immunohistochemistry Reveals Myeloid-Inflamed Tumor-Immune Complexity Associated with Poor Prognosis. *Cell Rep.* **19**, 203–

217 (2017).

232. Gerdes, M. J. *et al.* Highly multiplexed single-cell analysis of formalinfixed, paraffin-embedded cancer tissue. *Proc. Natl. Acad. Sci. U. S. A.* **110**, 11982–11987 (2013).

233. Roper, J. *et al.* In vivo genome editing and organoid transplantation models of colorectal cancer and metastasis. **35**, 569–576 (2017).

234. Kotecha, N., Krutzik, P. O. & Irish, J. M. Web-based analysis and publication of flow cytometry experiments. *Current Protocols in Cytometry* **Chapter 10**, (2010).

235. McKinley, E. T. *et al.* Machine and deep learning single-cell segmentation and quantification of multi-dimensional tissue images. *bioRxiv* 790162 (2019). doi:10.1101/790162

236. Bates, D., Mächler, M., Bolker, B. M. & Walker, S. C. Fitting linear mixed-effects models using lme4. *J. Stat. Softw.* (2015). doi:10.18637/jss.v067.i01

237. Lenth, R., Singmann, H., Love, J., Buerkner, P. & Herve, M. Package ' emmeans '. *R package version 1.15-15* (2020).

238. MATLAB. MATLAB 2020. *The MathWorks Inc.* (2021).

239. Radtke, F. & Clevers, H. Self-renewal and cancer of the gut: Two sides of a coin. *Science (80-.).* **307**, 1904–1909 (2005).

240. Schuijers, J. & Clevers, H. Adult mammalian stem cells: the role of Wnt, Lgr5 and R-spondins. *EMBO J.* **31**, 2685–2696 (2012).

241. Preston, S. L. *et al.* Bottom-up Histogenesis of Colorectal Adenomas. *Cancer Res.* **63**, 3819 LP – 3825 (2003).

242. Shih, I.-M. *et al.* Top-down morphogenesis of colorectal tumors. *Proc. Natl. Acad. Sci.* **98**, 2640–2645 (2001).

243. Goldenring, J. R. Pyloric metaplasia, pseudopyloric metaplasia, ulcer-associated cell lineage and spasmolytic polypeptide-expressing metaplasia: reparative lineages in the gastrointestinal mucosa. *J. Pathol.* **245**, 132 (2018).

244. Nam, K. T. *et al.* Mature chief cells are cryptic progenitors for metaplasia in the stomach. *Gastroenterology* (2010). doi:10.1053/j.gastro.2010.09.005

245. Schmidt, P. H. *et al.* Identification of a metaplastic cell lineage associated with human gastric adenocarcinoma. *Lab. Investig.* (1999).

246. Means, A. L. *et al.* Pancreatic epithelial plasticity mediated by acinar cell transdifferentiation and generation of nestin-positive intermediates. *Development* (2005). doi:10.1242/dev.01925

247. Meditskou, S., Grekou, A., Toskas, A., Papamitsou, T. & Miliaras, D. Pyloric and foveolar type metaplasia are important diagnostic features in crohn's disease that are frequently missed in routine pathology. *Histol. Histopathol.* (2020). doi:10.14670/HH-18-167

248. Thorsvik, S. *et al.* Ulcer-associated cell lineage expresses genes involved in regeneration and is hallmarked by high neutrophil gelatinase-associated lipocalin (NGAL) levels. *J. Pathol.* (2019). doi:10.1002/path.5258

249. Wright, N. A., Pike, C. & Elia, G. Induction of a novel epidermal growth factor-secreting cell lineage by mucosal ulceration in human gastrointestinal stem cells. *Nature* (1990). doi:10.1038/343082a0

250. Buczacki, S. J. A. *et al.* Intestinal label-retaining cells are secretory precursors expressing lgr5. *Nature* (2013). doi:10.1038/nature11965

251. Van Es, J. H. *et al.* Dll1 + secretory progenitor cells revert to stem cells upon crypt damage. *Nat. Cell Biol.* (2012). doi:10.1038/ncb2581

252. Schonhoff, S. E., Giel-Moloney, M. & Leiter, A. B. Neurogenin 3-expressing progenitor cells in the gastrointestinal tract differentiate into both endocrine and non-endocrine cell types. *Dev. Biol.* (2004). doi:10.1016/j.ydbio.2004.03.013

253. Tetteh, P. W. *et al.* Replacement of Lost Lgr5-Positive Stem Cells through Plasticity of Their Enterocyte-Lineage Daughters. *Cell Stem Cell* (2016). doi:10.1016/j.stem.2016.01.001

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

255. Bommi, P. V. *et al.* The Transcriptomic Landscape of Mismatch Repair-Deficient Intestinal Stem Cells. *Cancer Res.* **81**, 2760–2773 (2021).

256. Leach, J. D. G. *et al.* Oncogenic BRAF, unrestrained by TGFβ-receptor signalling, drives right-sided colonic tumorigenesis. *Nat. Commun. 2021 121* **12**, 1–15 (2021).

257. Tao, Y. *et al.* Aging-like Spontaneous Epigenetic Silencing Facilitates Wnt Activation, Stemness, and BrafV600E-Induced Tumorigenesis. *Cancer Cell* **35**, 315-328.e6 (2019).

258. Lieu, C. H. *et al.* Comprehensive Genomic Landscapes in Early and Later Onset Colorectal Cancer. *Clin. Cancer Res.* **25**, 5852–5858 (2019).

259. Dejea, C. M. *et al.* Microbiota organization is a distinct feature of proximal colorectal cancers. *Proc. Natl. Acad. Sci.* **111**, 18321–18326 (2014).

260. Kim, J. H. & Kang, G. H. Evolving pathologic concepts of serrated lesions of the colorectum. *J. Pathol. Transl. Med.* **54**, 276–289 (2020).

261. Knoop, K. A., McDonald, K. G., McCrate, S., McDole, J. R. & Newberry, R. D. Microbial sensing by goblet cells controls immune surveillance of luminal antigens in the colon. *Mucosal Immunol. 2015 81* **8**, 198–210 (2014).

262. Tallerico, R. *et al.* Human NK Cells Selective Targeting of Colon Cancer–Initiating Cells: A Role for Natural Cytotoxicity Receptors and MHC Class I Molecules. *J. Immunol.* **190**, 2381–2390 (2013).

263. Volonté, A. *et al.* Cancer-Initiating Cells from Colorectal Cancer Patients Escape from T Cell–Mediated Immunosurveillance In Vitro through Membrane-Bound IL-4. *J. Immunol.* **192**, 523–532 (2014).

264. Hofstadter 1945-, D. R. *Gödel, Escher, Bach : an eternal golden braid*. (New York : Basic Books, ©1979.).