Discovery of microenvironment drivers of cell states, plasticity and drug response

by

Andrew Warren Navia

B.S., Emory University (2015)

Submitted to the Department of Chemistry
in Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

May 2022

Signature of Author. . . . . .
Department of Chemistry
March 17, 2022

Certified by. . . . . .
Alex K. Shalek
Associate Professor of Chemistry
Thesis Supervisor

Accepted by. . . . . .
Adam Willard
Associate Professor of Chemistry
Graduate Officer

This doctoral thesis has been examined by a committee of the
Department of Chemistry as follows:

Laura L. Kiessling. . . . . .
Thesis Committee Chair
Novartis Professor of Chemistry

Alex K. Shalek. . . . .
Thesis Supervisor
Associate Professor of Chemistry

Scott Manalis... . . . . .
Thesis Committee Member
Andrew and Erna Viterbi Professor of Biological Engineering

Discovery of microenvironment drivers of cell states, plasticity and drug response

by

Andrew Warren Navia

Submitted to the Department of Chemistry
On March 17, 2022 in Partial Fulfillment of the
Requirements of the Degree of Doctor of Philosophy in Chemistry

Abstract

Cell state can be influenced by both intrinsic and extrinsic factors with functional consequences. Illustratively, in cancer, intrinsic genome level alterations or extrinsic microenvironmental immune cell activity can drive tumorigenesis. Similarly, in viral infections like COVID-19, the responses of infected and uninfected cells can impact clinical course. The recent emergence of single-cell genomic technologies like single-cell RNA sequencing (scRNA-seq) now enable us to characterize systematically and comprehensively the roles of intrinsic versus extrinsic responses in driving disease sequalae. Here, we apply these technologies to identify tumor cell states and their microenvironmental dependences, as well as to define infected cells and their supportive peripheral cells. Further, we establish new model systems based on *in vivo* interactions to nominate potential therapeutic targets.

Specifically, in pancreatic cancer, we refine a previously established basal and classical phenotype dichotomy and build on it by describing an intermediate state with a distinct supportive microenvironment. By understanding *in vivo* secreted factors from tumor and peripheral cells, we more accurately recapitulate cell-cell interactions *ex vivo,* allowing us to establish RNA state specific models. These *ex vivo* models suggest tumor cell plasticity that may play a role evasion of therapeutic pressure. Collectively, this work uncovers novel cancer biology, improves modeling of said biology and nominates therapeutic targets informed by system level interactions. Meanwhile, in COVID-19, we identify cell types prone to infection and tie disease severity to intrinsic epithelial immune responses. We associate clinical course with distinct immune environments, with severe cases harboring inflammatory macrophage populations and equivalent or elevated viral RNA load. We also identify viral targets, nominate mechanisms of viral entry, and find immune response trends in patients with severe disease.

Thesis Supervisor: Alex K. Shalek

Title: Core Member, IMES; Associate Professor, Chemistry; and, Extramural Member of the Koch Institute

Acknowledgments

I would first like to thank Alex Shalek for his mentorship and friendship over the past five years. I have learned how to be a better scientist, collaborator, and mentor from him, and I am incredibly grateful to have spent my graduate career in his lab.

Thank you to my committee, Professor Laura Kiessling, whose advice through these projects has been invaluable and Professor Scott Manalis, who I was fortunate to collaborate with extensively.

Thank you to past and present members of the Shalek lab, particularly those who trained and mentored me. I could not have done the work I did without all of you and the unwavering support you showed me.

Thank you to the PDAC team, particularly Dr. Peter Winter, Dr. Sri Raghavan and Jennyfer Galvez-Reyes with whom I worked most closely.

Thank you to the COVID-19 team, particularly Dr. Carly Ziegler, Vince Miao and Josh Bromley with whom I worked most closely.

Thank you to Suman Bose, Ben Mead, Laurent Jaquinod, Tom Demmitt and Shaina Carroll for your contributions to the bead projects.

Thank you to my pre-MIT scientific and academic mentors, Professor John Snyder, Professor Dennis Liotta and Professor Jeff Rosensweig as well as Dr. Rhonda Moore, Dr. Madhuri Dasari and the rest of the Liotta Lab.

Thank you to my many friends from Lexington, particularly Noah Resnick, Adrian Tanner and Alex Yared, as well friends from Emory, ice hockey and beyond.

Thank you to my partner Capt. Dr. Casey Anthony. I feel so fortunate to be able to lean on you during graduate school and celebrate both our accomplishments. Thank you to your family as well, they have been incredibly supportive.

Lastly, thank you to my family. To my paternal grandparents who sacrificed everything to move to the US to give me this opportunity. To my maternal grandparents who established deep family ties and an incredible support system. To my aunts, uncles and cousins who show me unconditional love. To my brother who's work ethic I strive to match and who makes me laugh. To my mom, who's leadership and enthusiasm I try to emulate and to my dad who got me into science in the first place but gave me space to be independent and explore my own interests.

Table of Contents

Lay summary

Humans are composed of trillions of cells working together to maintain organ health. Each cell reacts to those around it to maintain homeostasis and perform macroscopic functions. For example, white blood cells like T helper cells can recognize foreign pathogens in the body. This ability is made more influential given these cells' ability recruit other immune cells and mount a coordinated immune response to the invasion. When this synergy is disrupted however, the body becomes sick. These disruptions can arise from within cells, intrinsically, or from an outside source. Internal diseases can arise from mutation like ones that cause cancer and autoimmune diseases. External diseases are caused by viruses, like COVID-19, bacteria, and fungi among others. These can cause significant changes to the cells they infect and those surrounding. Regardless of mechanism, diseased cell states must be studied and their drivers identified.

Individual cells are incredibly complex; a single cell can dramatically influence its surrounding and propagate disease. As such, measuring cells individually can be incredibly powerful, revealing disease states and response mechanisms by surrounding cells. Until recently these sorts of single-cell measurements were often biased, needing a predetermined panel of markers to screen for. Recently however, single-cell RNA sequencing (scRNA-seq) methods have empowered comprehensive, unbiased single cell measurements. Measuring RNA is useful because it has numerous functions in the body. mRNA specifically draws from sections of the universal DNA script to code for proteins, the machinery of cells. scRNA-seq methods enrich for mRNA in an unbiased manner and shed light on acute cellular decision-making. In this document, we use scRNA-seq tools to identify novel cell states, interactions, and modeling methods in pancreatic ductal adenocarcinoma (PDAC) and COVID-19 infection samples.

PDAC is a devastating form of cancer with a 5-year survival of about 7%. Many therapeutic innovations successful in other cancers, like immune and genotype directed therapies, have been unsuccessful in PDAC. As such, researchers have shifted focus from traditional DNA profiling of PDAC tumor cells to RNA state profiling. These studies have discovered two prognostically distinct RNA states: basal, the more dire, and classical. While these studies were groundbreaking, their results were limited. Our team further investigates these states, measuring RNA expression within single tumor cells and associating particular phenotypes with surrounding non-malignant support structures. Critically, we find a third RNA state that suggests these tumor cells are plastic and can potentially escape drug pressure by changing phenotype. To probe these cells further we refine organoid model systems to better represent a patient’s tumor. Here, we apply different small molecules to shift states in a controlled manner. Lastly, in these high-fidelity models, we identify drug classes more suited to treat each tumor state. Moving forward, we plan to continue identifying PDAC therapies based on RNA expression and believe RNA state may be a critical component of drug evasion in cancer broadly.

As COVID-19 shut down the Institute and the world, I shifted my focus to studying viral infection. COVID-19 causes respiratory symptoms that can be fatal. Given the severity of the disease, with hundreds of millions of infections to date, there were concerted efforts across the scientific and medical communities to understand SARS-CoV-2, the virus that causes COVID-19. Fortunately, past coronaviruses like SARS (2002) and MERS (2012) offers a glimpse into COVID-19 biology. Early studies in the Shalek Lab and beyond identify exploitable similarities between SARS, MERS and COVID-19. Specifically, we use knowledge of how the virus enters human cells to predict likely target cells and examine how other diseases such as HIV might intersect with COVID-19. While these studies were important, relying on old datasets is limiting; as such, we carefully began to collect data using scRNA-seq on infected and control patient samples. Our data identifies which cell types were actually prone to infection and monitors how surrounding cells compensate for infection. Critically, we identify trends in patients with severe and mild COVID-19 infection. Here, patients with mild COVID-19 symptoms tend to have more a robust intrinsic immune response to the virus within the epithelial cells that line the nose. Moving forward, we hope to apply these lessons to emergent viral strains and model our findings in systems that accurately recapitulate infection severity.

In sum, our research uniquely leverages scRNA-seq protocols to probe critically important disease areas. In PDAC and COVID-19, understanding not only the diseased cells, but the way the surrounding microenvironment responds, has shed light on confounding factors regarding drug evasion (PDAC) and clinical course (COVID-19). This research emphasizes the importance of high-resolution methods that probe systems in an unbiased manner to translate academic research into clinical findings.

Statement of Contributions

The work described here was performed in collaboration with fellow Shalek lab members and numerous external collaborators.

COVID-19 work described in Chapter 2 was performed by members of the Shalek, Glover, Horwitz, and Ordovas-Montanes Labs. I helped outline experimental procedures, with a focus on safety, and performed much of the scRNA-seq wet lab work in collaboration with Dr. Carly Ziegler, Vince Miao and Josh Bromley. Lastly, I helped with manuscript preparation.

PDAC work described in Chapter 3 was performed by members of the Shalek, Wolpin, Aguirre, and Hahn labs. I designed scRNA-seq experiments with Dr. Alex Shalek, Dr. Peter Winter, Dr. Srivatsan Raghavan and Jennyfer Galvez-Reyes and perform all scRNA-seq work in collaboration with Jennyfer Galvez-Reyes. Additionally, I performed portions of the data analysis focused on 1) exploring model fidelity in mice and 2) identifying PDAC phenotypes in similar tumor types like cholangiocarcinoma and head and neck cancer. Lastly, I helped with manuscript preparation and figure generation.

Bead work briefly described in Chapters 1 and 4 was done primarily in the Shalek Lab with Shaina Carroll, Claire O'Callaghan and Dr. Ben Mead with contributions from Dr. Suman Bose in the Anderson/Sharp/Langer Labs. Dr. Laurent Jaquinod from Biolytic Lab Performance provided helpful advice throughout the project. I performed all oligosynthesis, scRNA-seq, data analysis and FACS sorting.

Lastly, work briefly mentioned in chapter 4 was done in collaboration with the Weinstock and Murakami labs. I performed all Seq-Well experiments in collaboration with Jennyfer Galvez-Reyes, Nolawit Mulugeta, Haley Strouf and Alejandro Gupta. Additionally, I performed portions of the data analysis with Dr. Peter Winter, Alan DenAdel and Sachit Saksena. Lastly, I helped with manuscript preparation and figure generation.

Chapter 1: Introduction

1.1 Importance of Cells

Cells have evolved to specialize based on the organ and microenvironment that they exist in. Multicellular organisms thrive due to synergistic interactions that allow for macroscopic function and system homeostasis. Normally, cell function is well regulated, cell reproduction is controlled, and intercellular interactions are productive.

Disease arises from dysregulation of native homeostasis. This can be intrinsically driven by DNA mutations that lead to detrimental gain or loss of function. They can also be extrinsically driven by external insults such as viral, bacterial or fungal infection, toxic substances, and radiation, among others. The body can respond to these perturbations by sequestering and destroying affected cells *via* an immune response. In certain cases, however, an immune response intended to return the system to normalcy can exacerbate illness. To understand the difference between health and disease and what drives productive versus unproductive immune response, precise tools capable of comprehensive cellular measurements are needed.

1.2 Towards Comprehensive measurements

Cells are complex systems and, as such, identifying cell types and measuring their function can take many forms[1-5]. Historically, cellular measurements have required compromise; either by averaging cellular expression in an unbiased manner across a whole tissue, or by isolating single cells and probing for specific features of interest, introducing biases[5-8]. These compromises extended across multiple measurable analytes like DNA, RNA and protein. Measurements of DNA expression require an average over many cells in order to reach a statistically meaningful signal threshold. This is due to the limited amount of DNA present per cell[5].

Messenger RNA (mRNA) and protein are more plentiful in cells and can describe cell type and state. Measurements of both have been developed to elucidate the mechanisms driving specific cell function and cellular interactions. Until recently however, these measurements also required compromise. Bulk level measurements, capable of measuring large quantities of RNA or protein from a tissue of interest, can mask an individual cell's role in either normal or disease states. In contrast, targeted probes, capable of measuring RNA or protein expression in single cells, are biased towards predetermined markers[5-8]. Despite limitations, these tools have been invaluable in

defining human health and disease at the molecular level. Groundbreaking studies have utilized bulk RNA sequencing (RNA-seq) to profile cell types in tissues, finding previously unappreciated cell states. For example, in pancreatic ductal adenocarcinoma (PDAC) Moffitt *et. al.* and others defined prognostically distinct RNA states independent of tumor cell genotype[9-11]. While this pioneering research draws a clear connection between RNA state and clinical outcome, the molecular structures that support these states and interactions between tumor and non-tumor cells remain elusive. Indeed, in these PDAC studies, RNA-seq is limited in its ability to assign a measured gene to a particular cell or cell type due to the bulk nature of the measurements. Assigning gene expression to single cells can clarify symbiotic interactions between cells and potentially describe how to therapeutically intervene in disease causing states. Performing these experiments on single cells to comprehensively measure these mediators of cellular function was not possible until recently[12-16]. The development of single cell RNA sequencing (scRNA-seq) methods has fundamentally changed how we characterize the role of individual cells in normal and disease states through the more focused and comprehensive measurements made possible by this technology.

1.3 Comprehensive scRNA-seq measurements

Isolating and measuring cells individually by scRNA-seq significantly improves data resolution. Today, dissociated tissue can be measured without enriching for specific cells of interest based on predetermined markers[17-19]. This not only allows for *de novo* identification of cell states but empowers identification and characterization of surrounding cells as well. As such, scRNA-seq methods have developed rapidly, but largely follow a similar pipeline. Broadly, mRNA is initially enriched by using a poly-adenylation targeting bait to pull down expressed genes. The mRNA is next reverse transcribed into a more stable cDNA and then amplified via PCR before being sequenced[13,14,20,21].

The first scRNA-seq protocol was published in 2009 by Tang *et. al.,* who hand-picked individual mouse embryonic cells (blastomeres) and captured nearly twice as many genes from a single cell as compared to the output from entire datasets utilizing bulk methods[12]. Over the past decade scRNA-seq technologies have further improved throughput, while lowering costs, and reagent and cell input requirements. The next major technique developed was SMRT-seq (Switching Mechanism at the 5'end of RNA Templates), in which single cells are flow sorted for processing,

into 96 or 384 well plates filled with lysis buffer. This technique significantly increased cell throughput and began to automate scRNA-seq protocols[15]. Two years after SMRT-seq was developed it was further improved (SMRT-seq2), delivering better gene detection, stability, and cDNA yields[16].

To dramatically increase throughput, some new methods moved away from 96 and 384 well plates to microfluidic systems and custom fabricated pico-well chips. Fluidics based protocols like 10X and Drop-seq rely on microfluidic systems to isolate cells and pair them with barcoded RNA capture beads[13,22,23]. While fluidics-based protocols are relatively expensive, due to the cost of the required fluids and the dead volume assigned per cell, they provide a more streamlined and user-friendly means of sequencing single cells than is possible in a plate format. Seq-well, the primary method employed in Chapters 2 and 3 of this document, is significantly cheaper than the 10X and Drop-seq methods, since it uses gravity and size exclusion to isolate cells into pico-wells[14]. When sealed with a semipermeable membrane, cells can be lysed and their mRNA can be captured on barcoded mRNA capture beads.

Most of these high throughput methods rely on bead-based mRNA capture[13,14,20]. This strategy has been in place for a number of years and is an area of comparatively slow innovation. The oligo strands on these beads are often built around a plastic core, with oligos added sequentially using a dedicated oligosynthsizer[13]. This process is cumbersome and requires extremely high yield reactions as well as technical precision. Moreover, bead loss in all these bead-based scRNA-seq protocols is high after the mRNA capture step. In response, I led projects that improve mRNA capture and enrichment as well as bead retention prior to sequencing. This work focuses on four areas: 1) developing a synthesis quality control pipeline, 2) building beads from a magnetic scaffold, 3) diversifying capture baits beyond poly-A and 4) generally increasing RNA captured and barcode diversity. These efforts have been ongoing but were unfortunately halted during the pandemic.

As cell recovery and throughput has improved so too has the statistical power of scRNA-seq datasets. These data have been used to infer interactions, identify putative tumor cells, and have been used extensively to atlas tissues in health[18,24-26]. An improved understanding of healthy tissues provides a robust comparator to benchmark the effect of disease. Given the comprehensive nature of scRNA-seq measurements, studies on disease can now move beyond a sole focus on infected or mutated cells, to characterize the surrounding environment and its possible role in

exacerbating dysfunction. In this work, the team and I apply these tools and concepts to profile direct and indirect cellular response to pancreatic cancer and SARS-CoV-2.

1.4 Using scRNA-seq to identify SARS-CoV-2 viral targets

SARS-CoV-2 the virus responsible for COVID-19 has become a generation-defining public health emergency. In March of 2020, our lab, the Institute, and much of the world shut down to curb the spread of this emergent virus. Unlike previous coronavirus outbreaks in 2002 (SARS) and 2012 (MERS), SARS-CoV-2 has spread worldwide, infecting hundreds of millions. While discrepancies between SARS, MERS and SARS-CoV-2 exist, the biology of those earlier viruses has proven vital in our understanding of the life cycle of SARS-CoV-2[27-29]. In particular, our understanding of viral entry has been informed by the body of research done on coronaviruses over the past two decades.

The infectivity of a virus is multifaceted, but partially informed by the mechanism utilized by the virus to enter a cell. Early SARS-CoV-2 research mined older datasets on viral entry to further infer the biology of the nascent virus[30]. These studies showed both SARS-CoV-2 and SARS use the protein ACE2, normally a regulator of blood pressure, as a mediator of cell entry through an interaction with the viral spike protein. Despite these early findings, many critical questions remain unanswered and require SARS-Cov-2 specific datasets for more thorough investigation.

To best understand the lifecycle of SARS-CoV-2 we must accurately identify cellular targets of the virus. Viral RNA is distinct from human RNA, creating a high-resolution flag when using scRNA-seq methods. The RNA expression of infected cells can be determined, quantified, and compared against uninfected cells. Differences can point to perturbed cellular functions and can narrow our focus on mechanisms of viral entry, replication, cell morbidity and death. Entry co-receptors, viral proteins, or native proteins that empower viral replication can also be identified as potential therapeutic targets. However, focusing strictly on infected cells ignores the potentially devastating compensatory response of the uninfected surroundings.

scRNA-seq can also define the role of uninfected cells in the microenvironment. Identifying secreted factors and surface receptors on un-infected cells may suggest intercellular interactions between immune and infected cells to further clarify the body's response to infection. Capturing and studying these surrounding cells may shed light on the discrepancies between a productive and unproductive immune response, leading to better therapeutics.

scRNA-seq is uniquely capable of measuring the state and activity of many cells comprehensively and in an unbiased manner. This is critical when dealing with a novel virus, since it can identify cellular interactions, system level responses and molecular differences tied to clinical outcome. Chapter 2 of this document draws from patient samples to answer these questions and others.

1.5 Measuring malignant and surrounding cells in PDAC

Collectively, cancer continues to be a leading cause of death world-wide. While survival rates have extended tremendously in certain cancers, pancreatic ductal adenocarcinoma (PDAC) in particular remains a bleak prognosis.[31] Genotype defined stratification schemes that have been clinically relevant in classifying and treating other cancers have had limited utility in PDAC[32]. Moreover, the relative lack of tumor cells and complex, stromal microenvironment common in PDAC make research and treatment difficult.

However, over the past ten years research teams have identified prognostically relevant RNA markers that distinguish clinical course[9-11]. These studies have been groundbreaking but limited in their description and precision, due to the bulk RNA methods they relied on. scRNA-seq, specifically Seq-Well, should provide a clearer and more detailed picture of tumor cell expression and possible cell state plasticity. This may clarify the nature of tumor states and their dependencies, despite tumor cell sparsity. Though mRNA markers have not been explicitly used in the clinic, they may yet provide a new lens with which to better understand and treat disease.

Microenvironmental heterogeneity has been another hurdle in profiling PDAC tumors. These tumors are characterized by dense stromal invasion, which might be involved in limiting therapeutic diffusion throughout the tumor[33]. Further, cutting edge immune therapies that have successfully directed an immune response against other tumor types, have shown little effect in PDAC. Some studies even show evidence that immune invasion can be counterproductive in treating these tumors[34]. This counterproductive immune response, along with diverse non-malignant populations justify deep profiling of tumor adjacent cells.

While RNA-base states have been proposed as a PDAC benchmark, our understanding of state dependencies and drivers have thus far been limited. Bulk RNA seq methods have blurred critical metrics of plasticity and intercellular interactions. Robust model system that accurately represent relevant phenotypes may shed light on tumor cell plasticity. However, while we can benchmark patient tumors *in vivo*, our ability to probe tumor cells *ex vivo* is blunted by biased model systems[34].

As such, there is a critical need for high fidelity model systems that recapitulate a patient's biology to test for plasticity, critical support structures and therapeutic efficacy. Experiments pursuing these goals and others are examined in Chapter 3.

1.6 Summary

scRNA-seq technologies have significantly changed the way scientists and clinicians measure causes of disease driven dysregulation. Fortunately, while biological applications continue to expand, technological innovation progresses as well[13,14,20,23-25]. Emergent protocols are continuing to improve throughput and sensitivity while lowering costs. The next major development may marry imaging— recording a cell's spatial coordinates—with scRNA-seq[21,35]. While these protocols and their associated computational tools are in their infancy, we plan to employ some of them in our research pipelines moving forward.

This thesis highlights both clinical translation and rapid exploratory research empowered by scRNA-seq while proposing future experiments that may encourage the use of the RNA state as a more clinically actionable biomarker. In summary, scRNA-seq is a critical tool in translating academic discovery into advancements in personalized medicine.

1.7 References

1. Emrich, S. J., Barbazuk, W. B., Li, L., & Schnable, P. S. (2007). Gene discovery and annotation using LCM-454 transcriptome sequencing. *Genome research*, *17*(1), 69–73.
2. Lister, R., O'Malley, R. C., Tonti-Filippini, J., Gregory, B. D., Berry, C. C., Millar, A. H., & Ecker, J. R. (2008). Highly integrated single-base resolution maps of the epigenome in Arabidopsis. *Cell*, *133*(3), 523–536.
3. Li Y. (2021). Modern epigenetics methods in biological research. *Methods (San Diego, Calif.)*, *187*, 104–113.
4. Drissi, R., Dubois, M. L., & Boisvert, F. M. (2013). Proteomics methods for subcellular proteome analysis. *The FEBS journal*, *280*(22), 5626–5634.
5. Mardis E. R. (2008). Next-generation DNA sequencing methods. *Annual review of genomics and human genetics*, *9*, 387–402.
6. Laerum, O. D., & Farsund, T. (1981). Clinical application of flow cytometry: a review. *Cytometry*, *2*(1), 1–13.

7. Schulze, A., & Downward, J. (2001). Navigating gene expression using microarrays--a technology review. *Nature cell biology*, *3*(8), E190–E195.
8. Marguerat, S., & Bähler, J. (2010). RNA-seq: from technology to biology. *Cellular and molecular life sciences : CMLS*, *67*(4), 569–579.
9. Moffitt, R.A., Marayati, R., Flate, E.L., Volmar, K.E., Loeza, S.G., Hoadley, K.A., Rashid, N.U., Williams, L.A., Eaton, S.C., Chung, A.H.*, et al.* (2015). Virtual microdissection identifies distinct tumor- and stroma-specific subtypes of pancreatic ductal adenocarcinoma. Nat Genet *47*, 1168-1178.
10. Collisson, E.A., Sadanandam, A., Olson, P., Gibb, W.J., Truitt, M., Gu, S., Cooc, J., Weinkle, J., Kim, G.E., Jakkula, L.*, et al.* (2011). Subtypes of pancreatic ductal adenocarcinoma and their differing responses to therapy. Nat Med *17*, 500-503.
11. Bailey, P., Chang, D.K., Nones, K., Johns, A.L., Patch, A.M., Gingras, M.C., Miller, D.K., Christ, A.N., Bruxner, T.J., Quinn, M.C.*, et al.* (2016). Genomic analyses identify molecular subtypes of pancreatic cancer. Nature *531*, 47-52.
12. Tang, F., Barbacioru, C., Wang, Y., Nordman, E., Lee, C., Xu, N., Wang, X., Bodeau, J., Tuch, B. B., Siddiqui, A., Lao, K., & Surani, M. A. (2009). mRNA-Seq whole-transcriptome analysis of a single cell. *Nature methods*, *6*(5), 377–382.
13. Macosko, E. Z., Basu, A., Satija, R., Nemesh, J., Shekhar, K., Goldman, M., Tirosh, I., Bialas, A. R., Kamitaki, N., Martersteck, E. M., Trombetta, J. J., Weitz, D. A., Sanes, J. R., Shalek, A. K., Regev, A., & McCarroll, S. A. (2015). Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets. *Cell*, *161*(5), 1202–1214.
14. Gierahn, T.M., Wadsworth, M.H., 2nd, Hughes, T.K., Bryson, B.D., Butler, A., Satija, R., Fortune, S., Love, J.C., and Shalek, A.K. (2017). Seq-Well: portable, low-cost RNA sequencing of single cells at high throughput. Nat Methods *14*, 395-398.
15. Ramsköld, D., Luo, S., Wang, Y. C., Li, R., Deng, Q., Faridani, O. R., Daniels, G. A., Khrebtukova, I., Loring, J. F., Laurent, L. C., Schroth, G. P., & Sandberg, R. (2012). Full-length mRNA-Seq from single-cell levels of RNA and individual circulating tumor cells. *Nature biotechnology*, *30*(8), 777–782.
16. Picelli, S., Björklund, Å. K., Faridani, O. R., Sagasser, S., Winberg, G., & Sandberg, R. (2013). Smart-seq2 for sensitive full-length transcriptome profiling in single cells. *Nature methods*, *10*(11), 1096–1098.

17. Prakadan, S. M., Alvarez-Breckenridge, C. A., Markson, S. C., Kim, A. E., Klein, R. H., Nayyar, N., Navia, A. W., Kuter, B. M., Kolb, K. E., Bihun, I., Mora, J. L., Bertalan, M. S., Shaw, B., White, M., Kaplan, A., Stocking, J. H., Wadsworth, M. H., 2nd, Lee, E. Q., Chukwueke, U., Wang, N., … Shalek, A. K. (2021). Genomic and transcriptomic correlates of immunotherapy response within the tumor microenvironment of leptomeningeal metastases. *Nature communications*, *12*(1), 5955.

18. Delorey, T. M., Ziegler, C., Heimberg, G., Normand, R., Yang, Y., Segerstolpe, Å., Abbondanza, D., Fleming, S. J., Subramanian, A., Montoro, D. T., Jagadeesh, K. A., Dey, K. K., Sen, P., Slyper, M., Pita-Juárez, Y. H., Phillips, D., Biermann, J., Bloom-Ackermann, Z., Barkas, N., Ganna, A., … Regev, A. (2021). COVID-19 tissue atlases reveal SARS-CoV-2 pathology and cellular targets. *Nature*, *595*(7865), 107–113.

19. Kazer, S. W., Aicher, T. P., Muema, D. M., Carroll, S. L., Ordovas-Montanes, J., Miao, V. N., Tu, A. A., Ziegler, C., Nyquist, S. K., Wong, E. B., Ismail, N., Dong, M., Moodley, A., Berger, B., Love, J. C., Dong, K. L., Leslie, A., Ndhlovu, Z. M., Ndung'u, T., Walker, B. D., … Shalek, A. K. (2020). Integrated single-cell analysis of multicellular immune dynamics during hyperacute HIV-1 infection. *Nature medicine*, *26*(4), 511–518.

20. Han, X., Wang, R., Zhou, Y., Fei, L., Sun, H., Lai, S., Saadatpour, A., Zhou, Z., Chen, H., Ye, F., Huang, D., Xu, Y., Huang, W., Jiang, M., Jiang, X., Mao, J., Chen, Y., Lu, C., Xie, J., Fang, Q., … Guo, G. (2018). Mapping the Mouse Cell Atlas by Microwell-Seq. *Cell*, *172*(5), 1091–1107.e17.

21. Stickels, R. R., Murray, E., Kumar, P., Li, J., Marshall, J. L., Di Bella, D. J., Arlotta, P., Macosko, E. Z., & Chen, F. (2021). Highly sensitive spatial transcriptomics at near-cellular resolution with Slide-seqV2. *Nature biotechnology*, *39*(3), 313–319.

22. Zheng, G. X., Terry, J. M., Belgrader, P., Ryvkin, P., Bent, Z. W., Wilson, R., Ziraldo, S. B., Wheeler, T. D., McDermott, G. P., Zhu, J., Gregory, M. T., Shuga, J., Montesclaros, L., Underwood, J. G., Masquelier, D. A., Nishimura, S. Y., Schnall-Levin, M., Wyatt, P. W., Hindson, C. M., Bharadwaj, R., … Bielas, J. H. (2017). Massively parallel digital transcriptional profiling of single cells. *Nature communications*, *8*, 14049.

23. Klein, A. M., Mazutis, L., Akartuna, I., Tallapragada, N., Veres, A., Li, V., Peshkin, L., Weitz, D. A., & Kirschner, M. W. (2015). Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. *Cell*, *161*(5), 1187–1201.

24. Jin, S., Guerrero-Juarez, C. F., Zhang, L., Chang, I., Ramos, R., Kuan, C. H., Myung, P., Plikus, M. V., & Nie, Q. (2021). Inference and analysis of cell-cell communication using CellChat. *Nature communications*, *12*(1), 1088.

25. inferCNV of the Trinity CTAT Project. https://github.com/broadinstitute/inferCNV
26. Wagner, J., Rapsomaniki, M. A., Chevrier, S., Anzeneder, T., Langwieder, C., Dykgers, A., Rees, M., Ramaswamy, A., Muenst, S., Soysal, S. D., Jacobs, A., Windhager, J., Silina, K., van den Broek, M., Dedes, K. J., Rodríguez Martínez, M., Weber, W. P., & Bodenmiller, B. (2019). A Single-Cell Atlas of the Tumor and Immune Ecosystem of Human Breast Cancer. *Cell*, *177*(5), 1330–1345.e18.
27. van Doremalen, N., Bushmaker, T., Morris, D. H., Holbrook, M. G., Gamble, A., Williamson, B. N., Tamin, A., Harcourt, J. L., Thornburg, N. J., Gerber, S. I., Lloyd-Smith, J. O., de Wit, E., & Munster, V. J. (2020). Aerosol and Surface Stability of SARS-CoV-2 as Compared with SARS-CoV-1. *The New England journal of medicine*, *382*(16), 1564–1567.
28. Cevik, M., Tate, M., Lloyd, O., Maraolo, A. E., Schafers, J., & Ho, A. (2021). SARS-CoV-2, SARS-CoV, and MERS-CoV viral load dynamics, duration of viral shedding, and infectiousness: a systematic review and meta-analysis. *The Lancet. Microbe*, *2*(1), e13–e22.
29. Hatmal, M. M., Alshaer, W., Al-Hatamleh, M., Hatmal, M., Smadi, O., Taha, M. O., Oweida, A. J., Boer, J. C., Mohamud, R., & Plebanski, M. (2020). Comprehensive Structural and Molecular Comparison of Spike Proteins of SARS-CoV-2, SARS-CoV and MERS-CoV, and Their Interactions with ACE2. *Cells*, *9*(12), 2638.
30. Ziegler, C., Allon, S. J., Nyquist, S. K., Mbano, I. M., Miao, V. N., Tzouanas, C. N., Cao, Y., Yousif, A. S., Bals, J., Hauser, B. M., Feldman, J., Muus, C., Wadsworth, M. H., 2nd, Kazer, S. W., Hughes, T. K., Doran, B., Gatter, G. J., Vukovic, M., Taliaferro, F., Mead, B. E., … HCA Lung Biological Network (2020). SARS-CoV-2 Receptor ACE2 Is an Interferon-Stimulated Gene in Human Airway Epithelial Cells and Is Detected in Specific Cell Subsets across Tissues. *Cell*, *181*(5), 1016–1035.e19.
31. Blackford, A. L., Canto, M. I., Klein, A. P., Hruban, R. H., & Goggins, M. (2020). Recent Trends in the Incidence and Survival of Stage 1A Pancreatic Cancer: A Surveillance, Epidemiology, and End Results Analysis. *Journal of the National Cancer Institute*, *112*(11), 1162–1169.
32. Singh, R. R., & O'Reilly, E. M. (2020). New Treatment Strategies for Metastatic Pancreatic Ductal Adenocarcinoma. *Drugs*, *80*(7), 647–669.
33. Biffi, G., & Tuveson, D. A. (2021). Diversity and Biology of Cancer-Associated Fibroblasts. *Physiological reviews*, *101*(1), 147–176.
34. Raghavan, S., Winter, P. S., Navia, A. W., Williams, H. L., DenAdel, A., Lowder, K. E., Galvez-Reyes, J., Kalekar, R. L., Mulugeta, N., Kapner, K. S., Raghavan, M. S., Borah, A. A., Liu, N., Väyrynen, S. A., Costa, A. D., Ng, R., Wang, J., Hill, E. K., Ragon, D. Y., Brais, L. K., … Shalek,

A. K. (2021). Microenvironment drives cell state, plasticity, and drug response in pancreatic cancer. *Cell*, *184*(25), 6119–6137.e26.

35. Moffitt, J. R., Hao, J., Wang, G., Chen, K. H., Babcock, H. P., & Zhuang, X. (2016). High-throughput single-cell gene-expression profiling with multiplexed error-robust fluorescence in situ hybridization. *Proceedings of the National Academy of Sciences of the United States of America*, *113*(39), 11046–11051.

Chapter 2: Impaired local intrinsic immunity to SARS-CoV-2 infection in severe COVID-19

AUTHORS

Carly G. K. Ziegler*, Vincent N. Miao*, Anna H. Owings*, Andrew W. Navia*, Ying Tang*, Joshua D. Bromley*, Peter Lotfy, Meredith Sloan, Hannah Laird, Haley B. Williams, Micayla George, Riley Drake, Taylor Christian, Adam Parker, L. Campbell Behlen, Molly W. Burger, Yilianys Pride, Kenneth J. Wilson, Mohammad Hasan, George E. Abraham, Michal Senitko, Tanya O. Robinson, Alex K. Shalek#, Bruce H. Horwitz#, Sarah C. Glover#, Jose Ordovas-Montanes#

* these authors contributed equally

# these senior authors contributed equally

2.1 Abstract

Infection with SARS-CoV-2, the virus that causes COVID-19, can cause severe lower respiratory illness including pneumonia and acute respiratory distress syndrome, which can lead to profound morbidity and mortality. Many infected individuals are either asymptomatic or have isolated upper respiratory symptoms, suggesting that the upper airways represent the initial site of viral infection, and that some individuals are able to largely constrain viral pathology to the nasal and oropharyngeal tissues. Despite major advances in understanding peripheral correlates of immunity and pathogenesis in COVID-19, which cell types in the human nasopharynx are the primary targets of SARS-CoV-2 infection, and how infection influences the cellular organization of the respiratory epithelium remains incompletely understood. Here, we present a cohort of nasopharyngeal samples from individuals with COVID-19, representing a wide spectrum of disease states from ambulatory to critically ill, as well as healthy and intubated patients without COVID-19. Using standard nasopharyngeal swabs, we collected viable cells and performed single-cell RNA-sequencing (scRNA-seq), simultaneously profiling both host and viral RNA. We find that following infection with SARS-CoV-2, the upper respiratory epithelium undergoes massive expansion and diversification of secretory cells and preferential loss of mature ciliated cells following infection with SARS-CoV-2. Active repopulation of lost ciliated cells appears to occur through secretory cell differentiation via deuterosomal cell intermediates. Epithelial cells from participants with

mild/moderate COVID-19 show extensive induction of genes associated with anti-viral and type I interferon responses. In contrast, cells from participants with severe lower respiratory symptoms appear globally stunted in their anti-viral capacity, despite substantially higher local inflammatory myeloid populations and equivalent nasal viral loads. This suggests an essential role for intrinsic, local, epithelial immunity in curbing and constraining viral infection. Using a custom computational pipeline, we characterized cell-associated SARS-CoV-2 RNA and identified rare cells with RNA intermediates strongly suggestive of active replication. We found remarkable diversity and heterogeneity among SARS-CoV-2 RNA+ host cells, both within and across individuals, including developing/immature and interferon-responsive ciliated cells, *KRT13*+ "hillock"-like cells, and unique subsets of secretory, goblet, and squamous cells. Finally, SARS-CoV-2 RNA+ cells, as compared to uninfected bystanders, were enriched for genes involved in either the cell-intrinsic response (e.g., *MX1*, *IFITM3*, *EIF2AK2*) or susceptibility to infection (e.g., *CTSL*, *TMPRSS2*). Together, this work defines both protective and detrimental host responses to SARS-CoV-2, determines the direct viral targets of infection, and suggests that failed cell-intrinsic anti-viral epithelial immunity in the nasal mucosa may underlie the progression to severe COVID-19.

2.2 Introduction

The novel coronavirus clade SARS-CoV-2 emerged in late 2019 and has quickly led to one of the most devastating global pandemics in modern history. Similar to other successful respiratory viruses, high replication within the nasopharynx[1,2] and viral shedding by asymptomatic or presymptomatic individuals contributes to high transmissibility[3,4] and rapid community spread[5-7]. COVID-19, the disease caused by SARS-CoV-2 infection, occurs in a fraction of those infected by the virus, and carries profound morbidity and mortality. The clinical pictures of COVID-19 vary widely – from some individuals who experience few mild symptoms to some with prolonged and severe disease characterized by pneumonia, acute respiratory distress syndrome, and diverse systemic effects impacting a variety of other tissues[8,9]. To facilitate effective preventative and therapeutic strategies for COVID-19, differentiating the host protective mechanisms that support rapid viral clearance and limit disease severity from those that drive severe and fatal outcomes is essential.

Rapid mobilization of the scientific community and a commitment to open data sharing early in the COVID-19 pandemic enabled researchers across the globe to study SARS-CoV-2 and build initial models of disease pathogenesis[10-12]. By analogy to related human betacoronaviruses[13], we currently understand viral tropism and disease progression to begin with SARS-CoV-2 entry through the mouth or nares where it initially replicates within epithelial cells of the human nasopharynx, generating an upper respiratory infection over several days[14]. A subset of patients develop symptoms of lower respiratory infection associated with viral replication in the distal airways, where a combination of inflammatory immune responses and direct viral-mediated pathogenesis can lead to diffuse damage to distal airways, alveoli, and vasculature[15,16]. Recent studies have mapped the host immune profiles associated with different stages along the COVID-19 disease trajectory. Reproducible immune correlates of severe COVID-19 include prolonged detection of proinflammatory cytokines such as IL-6, TNFα, and IL-8, diminished type I and type III interferon, and marked lymphopenia, as well as evidence for immune exhaustion and abnormal myeloid populations[17-23]. These reports have relied on host inflammatory and immune signatures from the peripheral blood, which may only partially reflect the immune status within virally targeted tissues[24,25]. To date, no large-scale studies have directly addressed the impact of SARS-CoV-2 infection on the respiratory epithelium of the human upper airways, nor assessed how this may relate to aberrant immune or anti-viral signaling described in the periphery.

A question central to understanding SARS-CoV-2 induced disease pathology is the precise identity of the direct cellular targets of viral infection within human respiratory tissues. Early in the pandemic, multiple groups conducted meta-analyses of existing single-cell transcriptomic datasets from diverse host tissues to map SARS-CoV-2 tropism based on *ACE2* expression and co-expression of host proteases required for spike protein cleavage[26-30]. Across these studies, the most likely SARS-CoV-2 targeted cells within the oropharyngeal, nasal, and upper airway tissues include subsets of ciliated, secretory, and goblet cells, while type II pneumocytes represent the most likely targets within the lung parenchyma. Indeed, a study using primarily bronchoalveolar lavage samples from a small cohort of COVID-19 patients identified rare SARS-CoV-2 RNA-containing cells assigned to ciliated and secretory cell types[31]. Further work using human tissues at autopsy found infected ciliated cells lining the trachea and distal airways within the lungs[32-34]. However, the precise early targets for SARS-CoV-2 in the nasopharynx, as well as the scope of

potential host cells and the variance in viral tropism across patients and disease courses have yet to be defined. A clearer understanding of viral tropism, how the airway epithelium responds to infection, and the relationship to disease outcome may critically inform therapeutic or prophylactic strategies.

In addition to cellular tropism, we currently lack a clear understanding of the host factors responsible for susceptibility vs. resistance to viral infection. Researchers have employed a variety of *in* vitro systems to assess induction of anti-viral defenses following SARS-CoV-2 infection. Compared to other common respiratory viruses, SARS-CoV-2 appears to poorly elicit type I interferon responses in cultured human epithelial cells, and instead skews towards proinflammatory cytokine profiles, in line with observations from human peripheral studies [17,35,36]. To directly assay virally-targeted cell types or tissues *in vivo,* researchers have relied on emerging animal models, including non-human primates[37-39], hamsters[40,41], mice[42-45], and ferrets[46,47]. Animal models vary widely in the severity of SARS-CoV-2-driven disease and associated immunopathology, and incompletely reflect the diversity of viral infection outcomes and natural immune responses within the human population[48]. Indeed, recent work has identified enrichment of both inborn errors of type I interferon signaling and the presence of auto-antibodies directed against type I interferons among patients with severe COVID-19, providing potential explanations for failed or insufficient anti-viral immunity within a subset of severe cases, and further supporting the need for studies of human cohorts that represent the breadth of host-viral interactions[49-51].

Here, we present a comprehensive analysis of the cellular phenotypes of the nasal mucosa during SARS-CoV-2 infection. To achieve this, we developed tissue handling protocols that enabled high-quality single-cell RNA-sequencing (scRNA-seq) from frozen nasopharyngeal swabs collected from a large patient cohort (n = 59), and created a detailed map of epithelial cell diversity and co-resident mucosal immune populations. We found that SARS-CoV-2 infection leads to a dramatic loss of mature ciliated cells, which is associated with secretory cell expansion, differentiation, and the accumulation of deuterosomal cell intermediates – potentially involved in the compensatory repopulation of damaged ciliated epithelium. Severe COVID-19 is characterized by mucosal recruitment of highly inflammatory myeloid populations which represent the primary sources of tissue pro-inflammatory cytokines including *TNF*, *IL1B*, and *CXCL8*, while type I and type III

interferons remain undetectable within resident immune or nasal epithelial cell types. Further, we identified profound differences in the induction of innate anti-viral pathways, genes involved in antigen processing and presentation, the acute inflammatory response, and pathways elicited by type I interferon between participants with mild/moderate vs. severe COVID-19. Finally, using unbiased whole-transcriptomic amplification, we were able to map not only host cellular RNA, but also cell-associated SARS-CoV-2 RNA, allowing us to trace viral tropism to specific epithelial subsets and identify host pathways associated with susceptibility or resistance to viral infection. Together, we identify an intrinsic failure of anti-viral immunity among nasal epithelial cells responding to SARS-CoV-2 infection, which predicts progression to severe COVID-19.

2.3 Results

2.3.1 Defining cellular Diversity in the Human Nasopharynx

Nasopharyngeal swabs were collected from 59 individuals from the University of Mississippi Medical Center between April and September 2020. This cohort consisted of 38 individuals who had a positive SARS-CoV-2 PCR nasopharyngeal (NP) swab on the day of hospital presentation. A Control cohort consisted of 15 individuals who were asymptomatic and had a negative SARS-CoV-2 NP PCR, and 6 intubated individuals in the intensive care unit without a recent history of COVID-19 and negative SARS-CoV-2 NP PCR (**Table 2.1**, see Methods for full inclusion and exclusion criteria). For the purposes of this study a second NP swab was collected within 3 days of presentation. Using the 2020 World Health Organization (WHO) guidelines for stratification and classification of COVID-19 severity based on the level of maximum required respiratory support, 16 of the individuals were considered COVID-19 mild/moderate (WHO score 1-5) and 22 had severe COVID-19 (WHO score 6-8, see **Methods**, **Table 2.1, Supplementary Figures 2.1A, 2.1B**). Nasopharyngeal samples were obtained by a trained healthcare provider and rapidly cryopreserved to maintain cellular viability (**Figure 2.1A, Supplementary Figure 2.1C**). Swabs were later processed to recover single-cell suspensions (mean +/- SEM: 57,000 +/- 15,000 total cells recovered per swab), before generating single-cell transcriptomes using the Seq-Well S^3 [52,53].

Table 2.1. Participant characteristics

	Control (WHO score 0)	Intubated Control (WHO score 7-8)	COVID-19 m/m (WHO score 1-5)	COVID-19 severe (WHO score 6-8)	COVID-19 conv. (WHO score 0)
Case number	25.9% (15/58)	10.3% (6/58)	24.1% (14/58)	36.2 (21/58)	3.4% (2/58)
Age (years)					
Minimum	27	33	19	28	20
Median (IQR)	58 (16)	65.5 (31)	49.5 (17.8)	62 (13)	N/A
Maximum	73	71	69	84	57
Sex					
Female	60% (9/15)	16.7% (1/6)	42.9% (6/14)	47.6% (10/21)	50% (1/2)
Male	40% (6/15)	83.3% (5/6)	57.1% (8/14)	52.4% (11/21)	50% (1/2)
Ethnicity					
Hispanic	0% (0/15)	0% (0/6)	0% (0/14)	4.8% (1/21)	0% (0/2)
Not Hispanic	100% (15/15)	100% (6/6)	100% (14/14)	95.2% (20/21)	100% (2/2)
Race					
Black/African American	66.7% (10/15)	66.7% (4/6)	71.4% (10/14)	61.9% (13/21)	50% (1/2)
White	33.3% (5/15)	33.3% (2/6)	28.6% (4/14)	23.8% (5/21)	50% (1/2)
American Indian	0% (0/15)	0% (0/6)	0% (0/14)	14.3% (3/21)	0% (0/2)
BMI					
Median (IQR)	37.5 (14.4)	30.5 (18.1)	23.0 (11.6)	31.9 (14.2)	40.7
Pre-existing conditions					
Diabetes	40% (6/15)	33.3% (2/6)	28.6% (4/14)	71.4% (15/21)	0% (0/2)
Chronic kidney disease	6.7% (1/15)	0% (0/6)	7.1% (1/14)	19.0% (4/21)	0% (0/2)
Congestive heart failure	6.7% (1/15)	16.7% (1/6)	0% (0/14)	4.8% (1/21)	0% (0/2)
Lung disorder	6.7% (1/15)	16.7% (1/6)	28.6% (4/14)	38.1% (8/21)	0% (0/2)
Hypertension	86.7% (13/15)	50% (3/6)	42.9% (6/14)	81.0% (17/21)	0% (0/2)
IBD	13.3% (2/15)	0% (0/6)	0% (0/14)	0% (0/21)	50% (1/2)
Treatment					
Corticosteroids	N/A	33.3% (2/6)	42.9% (6/14)	66.7% (14/21)	N/A
Remdesivir	N/A	0% (0/6)	42.9% (6/14)	85.7% (18/21)	N/A
28-day mortality	0% (0/15)	33.3% (2/6)	0% (0/14)	76.2% (16/21)	0% (0/2)

m/m: mild/moderate conv: convalescent IQR: inter-quartile range BMI: body mass index IBD: inflammatory bowel disease

Among all COVID-19 and Control samples, we recovered 32,871 genes across 32,588 cells (following filtering and quality control), with an average recovery of 562 +/- 69 cells per swab (mean +/- SEM). We found roughly equivalent transcriptomic quality following uniform preprocessing steps between COVID-1ry between participants (**Supplementary Figures 2.1D, 2.1E**). Following dimensionality reduction and clustering approaches to resolve individual cell types and cell states, we annotated 18 clusters corresponding to distinct cell types across immune

and epithelial identities (**Figure 2.1B-E, Supplementary Table 2.1**). Consistent with the use of nasal swabs for cell collection, we did not recover stromal cell populations such as endothelial cells, fibroblasts, or pericytes, which were found in previous scRNA-seq datasets from nasal epithelial surgical samples[54-56]. Among epithelial cell types, we readily identified basal cells by their expression of canonical marker genes including *TP63, KRT15,* and *KRT5*, as well as mitotic basal cells based on the added expression of genes involved in the cell cycle such as *MKI67* and *TOP2A* (**Figure 2.1F**). We resolved large populations of both secretory cells and goblet cells, identified by expression of *KRT7*, *CXCL17, F3, AQP5,* and *CP*. Despite strong transcriptional similarity between secretory and goblet cells, we distinguished between both cell types based on

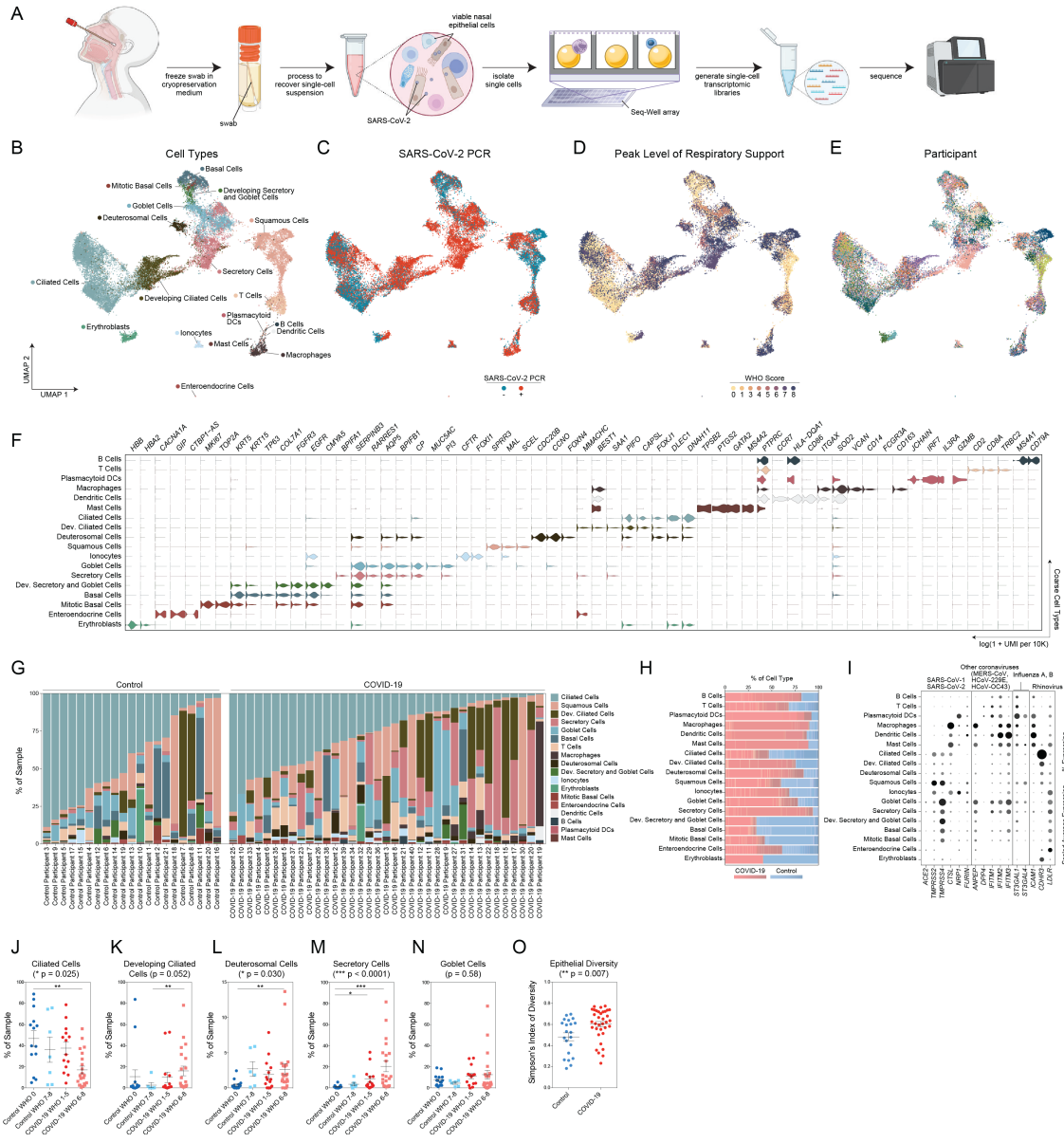


Figure 2.1: Cellular composition of nasopharyngeal swabs

(A) Schematic of method for viable cryopreservation of nasopharyngeal swabs, cellular isolation, and scRNA-seq using the Seq-Well S^3 platform (created with BioRender). **(B)**UMAP of 32,588 single-cell transcriptomes from all participants, colored by cell type (following iterative Louvain clustering). **(C)**UMAP as in **B**, colored by SARS-CoV-2 PCR status at time of swab. **(D)**UMAP as in **B**, colored by peak level of respiratory support (WHO COVID-19 severity scale). **(E)**UMAP as in **B**, colored by participant. **(F)**Violin plots of cluster marker genes (FDR < 0.01) for coarse cell type annotations (as in **B**). **(G)**Proportional abundance of coarse cell types by participant (ordered within each disease cohort by increasing Ciliated cell abundance). **(H)**Proportional abundance of participants by coarse cell types. Shades of red: COVID-19. Shades of blue: Control. **(I)**Expression of entry factors for SARS-CoV-2 and other common upper respiratory viruses. Dot size represents fraction of cell type (rows) expressing a given gene (columns). Dot hue represents scaled average expression. **(J)**Proportion of Ciliated Cells by sample. Statistical test above graph represents Kruskal-Wallis test results across all cohorts (following Bonferroni-correction). Statistical significance asterisks within box represent significant results from Dunn's post-hoc testing. * FDR-corrected p-value < 0.05, ** q < 0.01, *** q < 0.001. **(K)**Proportion of Developing Ciliated Cells by sample. **(L)**Proportion of Deuterosomal Cells by sample. **(M)**Proportion of Secretory Cells by sample. **(N)**Proportion of Goblet Cells by sample. **(O)**Simpson's Diversity index across epithelial cell types in COVID-19 vs. Control. Significance by Student's t-test.

expression of *MUC5AC*, which defines goblet cells, and *BPIFA1*, which we found primarily expressed within secretory cell types and diminished in *MUC5AC* high cells. We also designated a small population of cells "developing secretory and goblet cells" based on their lower expression of classic secretory/goblet cell genes, as well as persistent expression of some basal cell markers (e.g., persistent *COL7A1* and *DST* expression, but diminishing *KRT5*, *KRT15* expression). We also resolved a population of ionocytes, a recently-identified specialized subtype of secretory cell involved in regulating mucus viscosity within respiratory epithelia, defined by expression of *FOXI1, FOXI2*, and *CFTR*[57,58]. Squamous cells were identified by their expression of *SCEL*, as well as multiple *SPRR*- genes, and potentially derive from the squamous epithelium of the anterior nose or posterior pharynx. We also recovered a very small population of cells we term "enteroendocrine cells", based on unique expression of gastric inhibitory polypeptide (*GIP*), which is typically produced by intestinal and gastric enteroendocrine cells and *LGR5*, which classically marks stem cell populations in the gastrointestinal mucosa[59].

Ciliated cells were the most numerous epithelial cell type recovered in this dataset, defined by expression of transcription factor *FOXJ1* as well as numerous genes involved in the formation of cilia, e.g., *DLEC1, DNAH11,* and *CFAP43*. Similar to intermediate/developing cells of the

secretory and goblet lineage, we also identified two populations of precursor ciliated cells. One, termed "developing ciliated cells", which expressed canonical ciliated cell genes such as *FOXJ1*, *CAPSL*, and *PIFO* at lower levels than mature ciliated cells and lacked expression of cilia-forming genes. We also identified a cluster defined by expression of *DEUP1*, which is critical for centriole amplification as a precursor to cilium assembly. Together with co-expression of *CCNO, CDC20B, FOXN4,* and *HES6*, these cells match a recently-defined cell type termed deuterosomal cells, which represent a ciliated cell precursor cell type arising from secretory cell/goblet cell differentiation[55].

Immune cells represent a minority of recovered cells, yet we resolved multiple distinct clusters and cell types, representing major myeloid and lymphoid populations. Among lymphoid cells, we recovered T cells, identified by *CD3E, CD2,* and *TRBC2* expression, and B cells, identified by *MS4A1, CD79A,* and *CD79B* expression. Among myeloid cell types, we recovered a large population of macrophages (*CD14, FCGR3A, VCAN*), dendritic cells (*CCR7, CD86*), and plasmacytoid DCs (*IRF7, IL3RA*). Relative to true tissue-resident abundances, we under-recovered granulocyte populations, likely due to the intrinsic fragility of these cell types and the cryopreservation methods required in our sample pipeline. We recovered a very small population of mast cells, defined by expression of *GATA2, TPSB2,* and *PTGS2.* Among two samples, we recovered erythroblast-like cells, defined by expression of hemoglobin subunits including *HBB and HBA2*. With the exception of erythroblasts, each cell type was represented by cells from numerous participants, and from each participant we recovered a diversity of cell types and states, though the cellular composition was highly variable between distinct individuals (**Figure 2.1G, 2.1H**).

We directly tested whether cell types collected from nasal swabs following cryopreservation were representative of cellular composition extracted from a freshly swabbed nasal epithelium, or if certain cell types were lost during freezing (**Supplementary Figure 2.1F-2.1K**). Recovery of viable cells, technical metrics of single-cell library quality, and cellular proportions after clustering and analysis were all largely stable between matched fresh and cryopreserved swabs taken from the same individual. Importantly, no "new" cell types were recovered from the freshly processed samples (from healthy participants), suggesting that the on-swab cryopreservation technique

employed herein does not significantly alter the composition of cells available for downstream analysis.

We interrogated each cell type for the expression of host factors utilized by common respiratory viruses to facilitate cellular entry (**Figure 2.1I**)[29, 60-64]. We found *ACE2* expression highest among secretory cells and goblet cells, and to a lesser extent on ciliated cells, developing ciliated cells, deuterosomal cells, and squamous cells – suggesting these cells are likely targets for SARS-CoV-2 (and other betacoronaviruses that use ACE2 as their primary cellular entry factor). SARS-CoV-2 spike protein requires "priming" or cleavage by host proteases to enable membrane fusion and viral release into the cell – since early 2020, researchers have determined that proteases *TMPRSS2, TMPRSS4, CTSL,* and *FURIN* are capable of spike protein cleavage and are potentially critical for viral entry[60]. *TMPRSS2*, likely the principal host factor for SARS-CoV-2 S cleavage, is found in highest abundance on squamous cells, followed by modest expression on all other epithelial cell types. Similarly, *CTSL* (and other cathepsins) was found across diverse epithelial and myeloid cell types. *ANPEP* and *DPP4*, host receptors targeted by other human coronaviruses causing upper respiratory diseases, are found primarily on goblet cells and secretory cells[65,66]. As expected, *CDHR3*, the receptor utilized by Rhinovirus C, is found primarily on ciliated cells and developing ciliated cells[67].

Next, we grouped both SARS-CoV-2+ and SARS-CoV-2- participants by their level of respiratory support according to the WHO scoring system: Control WHO 0 (comprising healthy SARS-CoV-2 PCR negative participants, n = 15), Control WHO 7-8 (SARS-CoV-2 PCR negative, incubated participants treated in the ICU for non-COVID-19 diagnoses, n = 6), COVID-19 WHO 1-5 (SARS-CoV-2 PCR positive, mild/moderate disease, n = 14), and COVID-19 WHO 6-8 (SARS-CoV-2 PCR positive, intubated, severe disease, n = 21). We compared proportional cell type abundances across these four groups (**Figure 2.1J-2.1N**). We found that the abundance of ciliated cells (all, coarse annotation) was significantly impacted by group (Kruskal-Wallis test with Dunn's post-hoc testing, Bonferroni-corrected $p = 0.025$), and were significantly reduced among COVID-19 WHO 6-8 participants compared to healthy controls (mean +/- SEM 17.1 +/- 3.6 % of COVID-19 WHO 6-8 samples were ciliated cells, compared to 46.7 +/- 7.4 % of Control WHO 0, $p < 0.01$) (**Figure 2.1J**). Deuterosomal cells, which represent a developmental intermediate as secretory/goblet cells

differentiate into ciliated cells, were significantly increased among samples obtained from Control WHO 7-8, COVID-19 WHO 1-5, and COVID-19 WHO 6-8 samples, with the strongest increases observed among samples obtained from participants with severe COVID-19 compared to Control WHO 0 (**Figure 2.1L**). Likewise, developing ciliated cells were significantly increased among participants with severe COVID-19 (**Figure 2.1K**). The percentage of secretory cells was also dramatically increased among all COVID-19 participants compared to both the WHO 0 and WHO 7-8 control groups – 20.4 +/- 5.0% (mean +/- SEM) of all epithelial cells were secretory cells within severe COVID-19 participants, while mild/moderate COVID-19 participants contained 8.3 +/- 2.8% secretory cells, and on average, fewer than 4% of cells per participant were secretory among either Control WHO 0 and Control WHO 7-8 samples (**Figure 2.1M**). The average percentage of goblet cells was higher in both groups of participants with COVID-19 compared to controls, but this difference did not reach significance (**Figure 2.1N**). Intriguingly, expansion of secretory cells and loss of ciliated cells resulted in a net gain in epithelial diversity, calculated by Simpson's index which calculates the richness of the epithelial "ecosystem" (**Figure 2.1O**).

2.3.2 Epithelial Diversity and Remodeling Following SARS-CoV-2 Infection

Next, we sought to more completely delineate the diversity of epithelial cells through iterative clustering and sub-clustering among epithelial cell types (see **Methods**). This enabled us to divide the 10 "coarse" epithelial cell types into 25 "detailed" cell types/states (**Figures 2.2A-2.2E, Supplementary Figure 2.2A,** full differentiating gene lists for epithelial subtypes found in **Supplementary Table 2.1**). Among some cell types, we did not find additional within-type diversity, and thus the "coarse" annotations (**Figure 2.2A**) are equivalent to the "detailed" identities (**Figure 2.2D**). This applied to ionocytes, deuterosomal cells, developing secretory and goblet cells, basal cells, mitotic basal cells, and developing ciliated cells. We split goblet cells (coarse annotation) into 4 distinct detailed subtypes, each named by a representative defining marker or marker set. Likewise, secretory cells, squamous cells, and ciliated cells were all divided into multiple specialized subtypes. Some cellular subsets were similar to previously-described entities – including "*KRT24*^{high}*KRT13*^{high} secretory cells", which are highly similar to KRT13+ "hillock" cells, thought to be involved in airway epithelial responses to remodeling and inflammatory challenge[54,57]. Further, some cell types are defined by canonical cellular activation pathways, such as "interferon responsive" genes (e.g., *IFITM3*, *IFI6*, *MX1*) or "early response"

factors (e.g., *JUN*, *EGR1*, *FOS*). Finally, some cell types contained specialized transcriptomic profiles, which, to our knowledge, had not been previously characterized. This included a subset of squamous cells expressing markers classically associated with vascular endothelial cells including *VWF* and *VEGFA*, as well as secretory populations expressing high abundances of multiple inflammatory cytokines, such as "*BPIFA1*^{high}chemokine^{high} secretory cells" (chemokines include *CXCL8*, *CCL2*, *CXCL1*, and *CXCL3*) (**Figures 2.2D, 2.2E**).

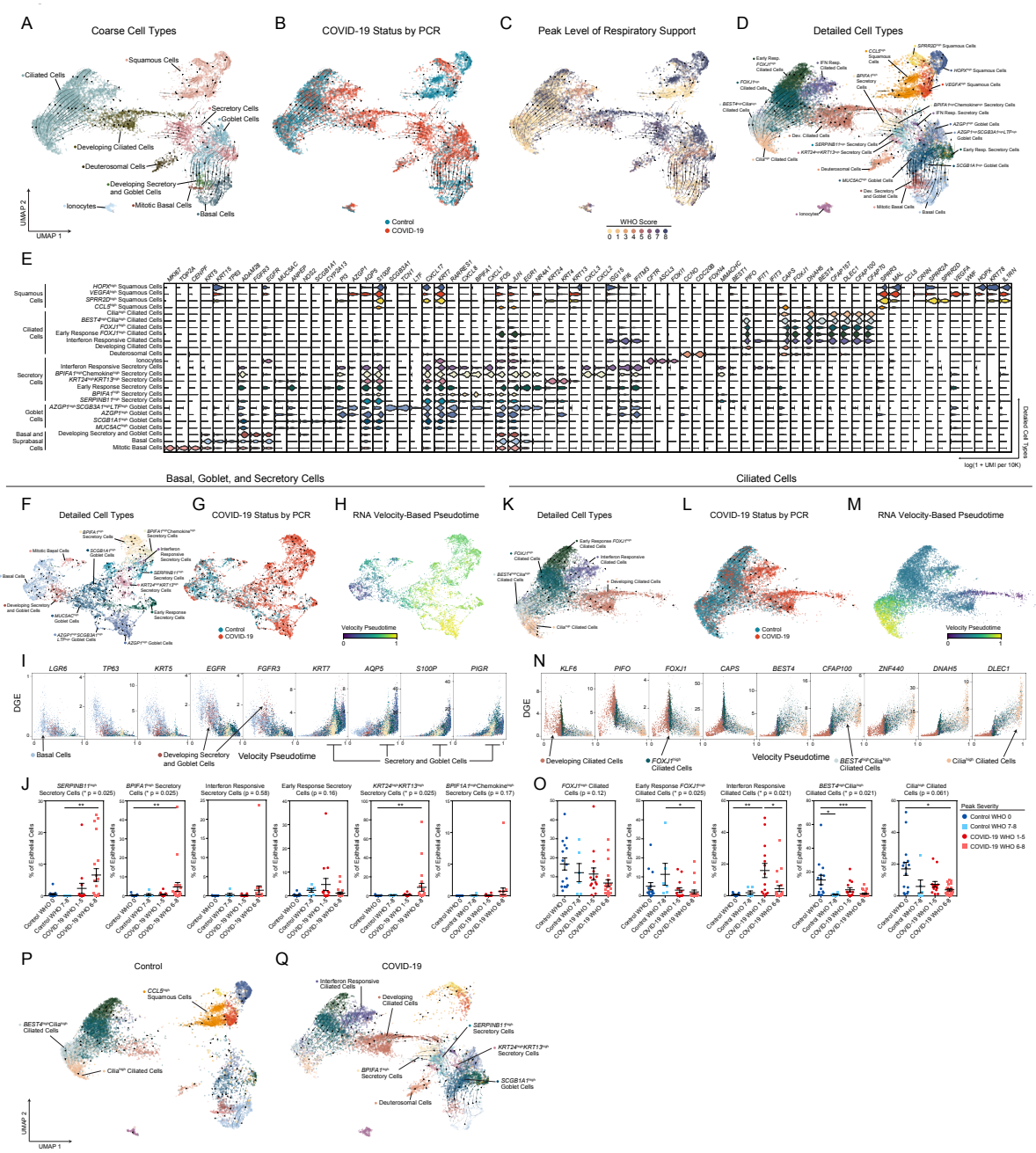


Figure 2.2: Altered epithelial cell composition and recovery in the nasopharynx during COVID-19
(A)UMAP of 28,948 epithelial cell types following re-clustering, colored by coarse cell types. Lines represent smoothed estimate of cellular differentiation trajectories (RNA velocity estimates via scVelo using intronic:exonic splice ratios). **(B)**UMAP as in **A**, colored by SARS-CoV-2 PCR status at time of swab. **(C)**UMAP as in **A**, colored by peak level of respiratory support (WHO illness severity scale). **(D)**UMAP as in **A**, colored by detailed cell annotations. **(E)**Violin plots of cluster marker genes (FDR < 0.01) for detailed epithelial cell type annotations (as in **D**). **(F)**UMAP of 9,209 Basal, Goblet, and Secretory Cells, following sub-clustering and resolution of detailed cell annotations. **(G)**UMAP of only Basal, Goblet, and Secretory Cells as in **F**, colored by SARS-CoV-2 PCR status at time of swab. **(H)**UMAP of only Basal, Goblet, and Secretory Cells as in **F**, colored by inferred velocity pseudotime (darker blue shades: precursor cells, lighter yellow shades: more terminally differentiated cell types). **(I)**Plot of gene expression by Basal, Goblet, and Secretory Cell velocity pseudotime for select genes. Points colored by detailed cell type annotations. **(J)**Proportion of Secretory Cell subtypes (detailed annotation) by sample, normalized to all epithelial cells. **(K)**UMAP of 13,913 Ciliated Cells, following sub-clustering and resolution of detailed cell annotations. **(L)**UMAP of Ciliated Cells as in **J**, colored by SARS-CoV-2 PCR status at time of swab. **(M)**UMAP of Ciliated Cells as in **J**, colored by inferred velocity pseudotime (darker blue shades: precursor cells, lighter yellow shades: more terminally differentiated cell types). **(N)**Plot of gene expression by Ciliated Cell velocity pseudotime for select genes (all significantly correlated with velocity expression. Points colored by detailed cell type annotations. **(O)**Proportion of Ciliated Cell subtypes (detailed annotation) by sample, normalized to all epithelial cells. **(P)**UMAP of 13,210 epithelial cells (using UMAP embedding from **A**) from SARS-CoV-2 PCR negative participants (Control). Lines represent smoothed estimate of cellular differentiation trajectories (via RNA velocity) calculated using only cells from Control participants. **(Q)**UMAP of 15,738 epithelial cells (using UMAP embedding from **A**) from SARS-CoV-2 PCR positive participants (COVID-19). Lines represent smoothed estimate of cellular differentiation trajectories (via RNA velocity) calculated using only cells from COVID-19 participants. Named cell types highlight those significantly altered between disease cohorts.

We again examined the epithelial subtypes for their expression of host entry factors which facilitate viral entry among common upper respiratory pathogens (**Supplementary Figure 2.2B**). *ACE2* was previously identified as highest among secretory, goblet, and ciliated cells[29,30] – here we observe substantial within-cell type heterogeneity in *ACE2* expression among each of these cell types. Notably, among goblet cells, *AZGP1*^{high} goblet cells express the highest abundance of *ACE2* mRNA, suggesting this cell type may be a preferential target for SARS-CoV-2 infection. Likewise, early response secretory cells, *KRT24*^{high}*KRT13*^{high} secretory cells, and interferon responsive secretory cells, all express elevated abundances of *ACE2*. Many other secretory and goblet cell types express detectable *ACE2*, but lower levels. Similarly, multiple detailed subsets of ciliated cells expressed *ACE2*, however cilia^{high} and *BEST4*^{high}cilia^{high} ciliated cells notably did not appear to contain detectable levels of *ACE2* mRNA.

To map the differentiation trajectories and lineage relationships between epithelial cell types, we analyzed single-cell RNA velocity (scVelo) across all epithelial cells[68,69]. RNA velocity analysis leverages the dynamic relationships between expression of unspliced (intron-containing) and spliced (exonic) RNA across thousands of variable genes, enabling 1) estimation of the directionality of transitions between distinct cells and cell types, and 2) identification of putative driver genes behind these transitions. Overlaying the UMAPs of cell type identities and associated metadata in **Figures 2.2A-2.2D**, vector fields (black lines and arrows) represent a smoothed estimate of cellular transitions based on RNA velocity. Globally, RNA velocity appropriately places basal cells and mitotic basal cells as the "root" or "origin" of cellular transitions, which then progress through the developing secretory and goblet cells to the secretory cells and goblet cells. Developing ciliated cells and ciliated cells are placed "later" in the differentiation trajectory, distal to development of both secretory and deuterosomal cells, which is consistent with current models where ciliated cells represent a terminally differentiated state and may arise from these precursor cell types[55]. By analyzing spliced and unspliced forms of representative markers underlying ciliated cell development, we can visualize the transition from secretory cells to deuterosomal cells to developing ciliated cells, and finally mature ciliated cells (**Supplementary Figure 2.2C**). Together, this analysis enables us to map the developmental relationships between major epithelial cell compartments discussed above, and connect the loss of "terminally differentiated" or "mature" cell types in COVID-19, e.g., ciliated cells, with the concurrent expansion of their apparent precursors: secretory, deuterosomal, and developing ciliated cells (**Figure 2.1J-2.1N**).

We next analyzed developmental transitions *among* detailed epithelial cell subtypes (as presented in **Figure 2.2D**) to better trace the relationships between finer-resolved subsets, and map alterations in cellular behavior and development during COVID-19. When considering only basal, goblet, and secretory cell subtypes, we found *TP63*, *KRT5,* and *LGR6* expression gradually decline across basal and developing secretory and goblet cells, while expression of secretory and goblet cell specific markers such as *KRT7* and *AQP5* progressively increase (**Figure 2.2F-2.2I**). The majority of secretory and goblet clusters are represented by cells from SARS-CoV-2+ individuals (as observed previously, **Figure 2.1K, 2.2G**), with significant expansion of *SERPINB11*^{high} secretory cells (which represent a "generic" or un-differentiated secretory subtype), *BPIFA1*^{high}

secretory cells, and *KRT24*^{high}*KRT13*^{high} secretory cells (which resemble KRT13+ "hillock" cells) among cells from individuals with severe COVID-19 (**Figure 2.2J**). Notably, transitions between detailed secretory and detailed goblet cells are substantially less linear than among the coarse cell types or as seen in ciliated cell subsets (discussed below). RNA velocity curves predict multiple routes for development between different secretory and goblet subtypes (**Figure 2.2F**), which suggests maintained capacity for differentiation and de-differentiation even among this "mature" cell type, and is consistent with the current understanding of respiratory secretory cell plasticity[70].

Ciliated cell subtypes were analyzed by their RNA velocity and pseudotemporal ordering in the same manner (**Figures 2.2K-2.2N**). The velocity pseudotime predicts progression from developing ciliated cells, to *FOXJ1*^{high} ciliated cells, to *BEST4*^{high}cilia^{high} ciliated cells, and terminating in cilia^{high} ciliated cells. (**Figure 2.2M**). Interferon responsive ciliated cells and early response *FOXJ1*^{high} ciliated cells represent phenotypic deviations from this ordered progression, and therefore appear collapsed/unresolved along this trajectory with the same pseudotime range as *FOXJ1*^{high} ciliated cells. Among COVID-19 participants, we observed decreased proportions of both cilia^{high} and *BEST4*^{high}cilia^{high} ciliated cells, two cell subsets which represent the most terminally differentiated ciliated cell subtypes (**Figure 2.2O**). This effect was particularly pronounced among individuals with severe disease, and suggests that the overall reduction in upper airway ciliated cells during COVID-19 (**Figure 2.1J)** preferentially affects terminally differentiated subsets, potentially due to delayed replenishing from secretory/deuterosomal precursors, or enhanced susceptibility to viral-mediated pathogenesis. Among individuals with mild/moderate COVID-19, we found a substantial increase in the proportion of interferon responsive ciliated cells – averaging 15.9% of all epithelial cells among mild/moderate COVID-19 participants, compared to < 1% among healthy controls (**Figure 2.2O**).

Finally, we directly mapped the developmental transitions among nasal epithelial cells within Control (**Figure 2.2P**), or COVID-19 participants only (**Figure 2.2Q**). Confirming our above analysis, cells from Control participants poorly populated the intermediate regions that bridge secretory and goblet cell types to mature ciliated cells. Conversely, regions annotated as multiple secretory cell subsets and developing ciliated cells were uniquely captured from COVID-19 participants. Together, our analysis defines both the cellular diversity among cells collected from

nasopharyngeal swabs, as well as the nuanced developmental relationships between epithelial cells of the upper airway. Further, we observe substantial expansion of immature/intermediate and specialized subtypes of secretory, goblet, and ciliated cells during COVID-19, presumably as a result of direct viral targeting and pathology, as well as part of the intrinsic capacity of the nasal epithelium to regenerate and repopulate following damage.

2.3.3 Alterations to Nasal Mucosal Immune Populations in COVID-19

As with epithelial cells, we further clustered and annotated detailed immune cell populations. Multiple cell types could not be further subdivided from their coarse annotation (**Figure 2.1B, Supplementary Figure 2.3A-2.3E**), including mast cells, plasmacytoid DCs, B cells, and dendritic cells. Among macrophages (coarse annotation), we resolved 5 distinct subtypes (**Supplementary Figure 2.3B**). *FFAR4*^{high} macrophages were defined by expression of *FFAR4*, *MRC1*, *CHIT1*, and *SIGLEC11*, as well as chemotactic factors including *CCL18*, *CCL15*, genes involved in leukotriene synthesis (*ALOX5, ALOX5AP, LTA4H*), and toll-like receptors *TLR8* and *TLR2* (**Supplementary Figure 2.3F**, full differentiating gene lists for immune subtypes found in **Supplementary Table 2.1**). Interferon responsive macrophages were distinguished by elevated expression of anti-viral genes such as *IFIT3, IFIT2, ISG15,* and *MX1*, akin to the epithelial subsets labeled "interferon responsive", along with *CXCL9*, *CXCL10*, *CXCL11*, which are likely indicative of IFNγ stimulation. *MSR1*^{high}*C1QB*^{high} macrophages are defined by cathepsin expression (*CTSD, CTSL, CTSB*) and elevated expression of complement (*C1QB, C1QA, C1QC*), and lipid binding proteins (*APOE, APOC,* and *NPC2*). The fourth "specialized" subtype of macrophage we found was termed "inflammatory macrophages", which uniquely expressed inflammatory cytokines such as *CCL3, CCL3L1, IL1B, CXCL2,* and *CXCL3*. The remaining "*ITGAX*^{high}" macrophages were distinguished from other immune cell types by *ITGAX*, *VCAN, PSAP, FTL, FTH1* and *CD163* (though these genes are shared by other specialized macrophages subsets). T cells were largely *CD69* and *CD8A* positive, consistent with a T resident memory-like phenotype, and we were not able to resolve a separate cluster of CD4 T cells. Two specialized subtypes of CD8 T cells were annotated from this dataset: one defined by exceptionally high expression of early response genes (*FOSB, NR4A2,* and *CCL5*), and the other termed "interferon responsive cytotoxic CD8 T cells", defined by granzyme and perforin expression (*GZMB, GZMA, GNLY, PRF1, GZMH*), anti-viral

genes (*ISG20, IFIT3, APOBEC3C, GBP5*) and genes associated with effector CD8 T cell function (*LAG3, IL2RB, IKZF3, TBX21*).

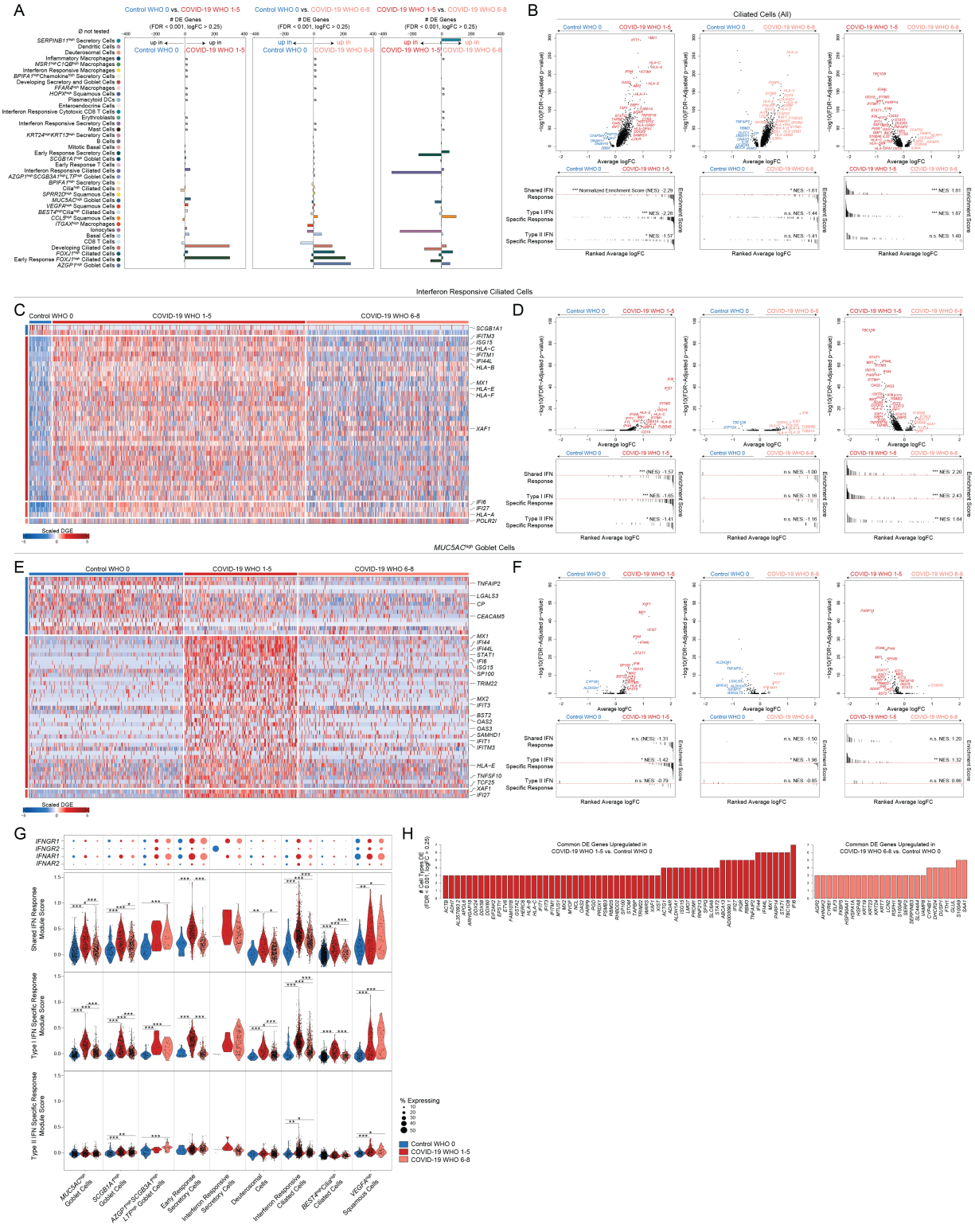


Figure 2.3: Cell-type specific and shared transcriptional responses to SARS-CoV-2 infection.
(A) Abundance of significant differentially expressed (DE) genes by detailed cell type between Control WHO 0 vs. COVID-19 WHO 1-5 samples (left), Control WHO 0 and COVID-19 WHO 6-8 samples (middle), COVID-19 WHO 1-5 and COVID-19 WHO 6-8 samples (right). Restricted to genes with FDR-corrected $p < 0.001$, log2 fold change > 0.25. ø = comparison not tested due to too few cells in one group. **(B)** Top: Volcano plots of average log fold change vs. $-\log_{10}$(FDR-adjusted p-value) for Ciliated cells (coarse annotation). Left: Control WHO 0 vs. COVID-19 WHO 1-5 (mild/moderate). Middle: Control WHO 0 vs. COVID-19 WHO 6-8 (severe). Right: COVID-19 WHO 1-5 (mild/moderate) vs. COVID-19 WHO 6-8 (severe). Horizontal red dashed line: FDR-adjusted p-value cutoff of 0.05 for significance. Bottom: gene set enrichment analysis plots across shared, type I interferon specific, and type II interferon specific stimulated genes. Genes are ranked by their average log fold change (FC) between each comparison. Black lines represent the ranked location of genes belonging to the annotated gene set. Bar height represents running enrichment score (NES: Normalized Enrichment Score). P-values following Bonferroni-correction: *** corrected $p < 0.001$, ** $p < 0.01$, * $p < 0.05$. **(C)** Heatmap of significantly DE genes between Interferon Responsive Ciliated Cells from different disease cohorts. **(D)** Top: Volcano plots related to **C**. Average log fold change vs. $-\log_{10}$(FDR-adjusted p-value) for Interferon Responsive Ciliated cells. Horizontal red dashed line: 0.05 cutoff for significance. Bottom: gene set enrichment analysis across shared, type I, and type II interferon stimulated genes. **(E)** Heatmap of significantly DE genes between *MUC5AC*^{high} Goblet Cells from different disease cohorts. **(F)** Top: Volcano plots related to **E**. Average log fold change vs. $-\log_{10}$(FDR-adjusted p-value) for *MUC5AC*^{high} Goblet Cells. Horizontal red dashed line: 0.05 cutoff for significance. Bottom: gene set enrichment analysis across shared, type I, and type II interferon stimulated genes. **(G)** Top: Dot plot of *IFNGR1/2* and *IFNAR1/2* gene expression by selected cell types. Bottom: Violin plots of gene module scores across selected cell types, split by Control WHO 0 (blue), COVID-19 WHO 1-5 (red), and COVID-19 WHO 6-8 (pink). Gene modules represent transcriptional responses of human basal cells from the nasal epithelium following *in vitro* treatment with IFNA or IFNG. Significance by Wilcoxon signed-rank test. P-values following Bonferroni-correction: * $p< 0.05$, ** $p < 0.01$, *** $p < 0.001$. **(H)** Common DE genes across detailed cell types. Left (red): genes upregulated in multiple cell types when comparing COVID-19 WHO 1-5 vs. Control WHO 0. Right (pink): genes upregulated in multiple cell types when comparing COVID-19 WHO 6-8 vs. Control WHO 0.

Among immune cells, macrophages were markedly increased relative to other immune cell types during severe COVID-19 (**Supplementary Figure 2.3G, 2.3H**). Multiple specialized myeloid cell types were uniquely detected and enriched among COVID-19 participants, albeit in a subset of participants, and biased to severe COVID-19 cases: *ITGAX*^{high} macrophages, *FFAR*4^{high} macrophages, inflammatory macrophages, and interferon responsive macrophages (**Supplementary Figure 2.3H**). Through rare, plasmacytoid DCs and mast cells were only recovered as > 1% of immune cells among COVID-19 participants. Somewhat surprisingly, T cells and T cell subtypes were not dramatically altered between disease cohorts. Finally, we assessed the correlation between distinct cell types across all participants. When samples from all disease cohorts were considered, we found that proportional abundance of dendritic cells, mast cells, and macrophages were highly-correlated with one another ($p < 0.01$), likely indicative of the coordinated recruitment of these immune subtypes during inflammation. Among detailed immune cell types, interferon responsive macrophages were highly correlated with interferon responsive cytotoxic CD8 T cells ($p < 0.01$), suggesting direct communication between *IFNG*-expressing tissue resident T cells and *CXCL9/10/11* expressing myeloid cells.

These analyses demonstrate how the epithelial and immune compartments are dramatically altered during COVID-19, likely reflecting both protective anti-viral and regenerative responses, as well as pathologic changes underlying progression to severe disease.

2.3.4 Cellular Behaviors Associated with COVID-19 Severity

Thus far, we have characterized how severe COVID-19 elicits major cell compositional changes within the nasopharyngeal mucosa, including expansion of the secretory cell/deuterosomal cell compartments associated with lost mature ciliated cells, and recruitment of highly inflammatory myeloid cells. Next, we examined how each individual cell type responds across the full spectrum of disease severity. Here, we analyzed pairwise comparisons between Control WHO 0, COVID-19 WHO 1-5 (mild/moderate), and COVID-19 WHO 6-8 (severe), and compared both high-level "coarse" cell types, and "detailed" cell subsets (**Figure 2.3A, Supplementary Figure 2.4A, Supplementary Tables 2.2-2.4**). Among all coarse cell types, the largest magnitude transcriptional changes (measured by the number of differentially expressed (DE) genes with FDR < 0.001, and log fold change > 0.25) were observed primarily within the epithelial compartment, most strikingly within ciliated cells, developing ciliated cells, secretory cells, goblet cells, and ionocytes (**Supplementary Figure 2.4A**). Among detailed cell types, we observed the largest transcriptional changes among *AZGP1*^{high} goblet cells, early response *FOXJ1*^{high} ciliated cells, *FOXJ1*^{high} ciliated cells, *MUC5AC*^{high} goblet cells, *SERPINB11*^{high} secretory cells, early response secretory cells, and interferon responsive ciliated cells. Broadly, major differences were observed in the *identity* of cell types with large transcriptional responses – with mild/moderate COVID-19 driving differences principally in *MUC5AC*^{high} goblet cells and ionocytes, while severe COVID-19 included major perturbations among basal cells, *AZGP1*^{high} goblet cells. Ciliated subsets were profoundly altered in both mild/moderate and severe COVID-19 compared to cells from Control WHO 0 participants. Finally, when we directly compared mild/moderate to severe COVID-19, multiple cell types showed robust transcriptional changes, most drastically among ciliated cell subtypes (interferon responsive ciliated cells, *FOXJ1*^{high} ciliated cells, early response *FOXJ1*^{high} ciliated cells, developing ciliated cells), ionocytes, *SERPINB11*^{high} secretory cells, early response secretory cells, and *AZGP1*^{high} goblet cells.

We next examined the specific DE genes among ciliated cells (all, coarse annotation) between each cohort (**Figure 2.3B**, **Supplementary Tables 2.2-2.4**). Compared to ciliated cells from Control WHO 0 participants, cells from both mild/moderate COVID-19 and severe COVID-19 robustly upregulated genes involved in the host response to virus, including *IFI27, IFIT1, IFI6, IFITM3,* and *GBP3*, and both cohorts induced expression of MHC-I and MHC-II genes (including *HLA-A, HLA-C, HLA-F, HLA-E, HLA-DRB1, HLA-DRA*) and other factors involved in antigen processing and presentation (**Supplementary Figures 2.4B, 2.4C**). Notably, large sets of interferon-responsive and anti-viral genes were exclusively induced among ciliated cells from COVID-19 WHO 1-5 participants when compared to Control WHO 0 participants. In a direct comparison of ciliated cells from mild/moderate to severe COVID-19, the cells from individuals with mild/moderate disease showed strong upregulation of diverse anti-viral factors, including *IFI44L, STAT1, IFITM1, MX1, IFITM3, OAS1, OAS2, OAS3, STAT2, TAP1, HLA-C, ADAR, XAF1, IRF1, CTSS, CTSB*, and many others (**Supplementary Figure 2.4C**). Ciliated cells from severe COVID-19 uniquely upregulated *IL5RA* and *NLRP1* (compared to both control and mild/moderate COVID-19). Together, these DE gene sets are suggestive of exposure to secreted inflammatory factors and type I/II/III interferons, as well as direct cellular sensing of viral products. Using previously published data from human nasal basal cells treated *in vitro* with either type I (IFNγ) or type II (IFNγ) interferon[30], we created gene sets that represented the "shared" gene responses to type I and type II interferon, and the cellular responses specific to either type (**Figure 2.3B**). Using gene set enrichment analysis, we tested whether the genes that discriminate ciliated cells from different groups (e.g., mild/moderate COVID-19 vs. severe COVID-19) imply exposure to specific interferon types. We found that ciliated cells in mild/moderate COVID-19 robustly induced type I interferon-specific gene signatures, both compared to cells from healthy controls, as well cells from severe COVID-19. Further, when compared to cells from healthy individuals, ciliated cells from individuals with severe COVID-19 did not significantly induce type I or type II interferon responsive genes, potentially underlying poor control of viral spread.

We next investigated whether these effects were observed among other cell types and subsets. Surprisingly, even *among* cells defined as "interferon responsive" ciliated cells, cells from mild/moderate COVID-19 participants expressed higher fold changes of interferon-responsive genes compared to cells from COVID-19 WHO 6-8 participants or Control WHO 0 (**Figures 2.3C,**

2.3D, Supplementary Tables 2.2-2.4). Other detailed epithelial cell types displayed a similar pattern: broad interferon-responsive genes (largely type I specific) were strongly upregulated among cells from mild/moderate COVID-19 participants, while cells from severe COVID-19 upregulated few shared markers with mild/moderate COVID-19 participants, and instead skewed towards inflammatory genes such as *S100A8* and *S100A9* instead of anti-viral factors (**Figures 2.3E-2.3H, Supplementary Figure 2.4D**). In some cases, cells from individuals with severe COVID-19 expressed levels of interferon responsive or anti-viral genes indistinguishable from healthy controls. Strongest induction of type I specific interferon responses among mild/moderate COVID-19 cases was observed in *MUC5AC*^{high} goblet cells, *SCGB1A1*^{high} goblet cells, early response secretory cells, deuterosomal cells, interferon responsive ciliated cells, and *BEST4*^{high}cilia^{high} ciliated cells (**Figure 2.3G**). Rare cell types from severe COVID-19 individuals induced comparable type I interferon responses to their mild/moderate counterparts, including *AZGP1*^{high}*SCGB3A1*^{high}*LTF*^{high} goblet cells, interferon responsive secretory cells, and *VEGFA*^{high} squamous cells. Expression of type II specific genes were globally blunted across all cell types from COVID-19 samples when compared to type I module scores (**Figure 2.3G, Supplementary Figures 2.3K, 2.4D**). Further, the absence of a transcriptional response to secreted interferon could not be explained by a lack of either interferon alpha receptor (*IFNAR1, IFNAR2*) or interferon gamma receptor (*IFNGR1, IFNGR2*) expression. Previous work has identified *ACE2*, the host receptor for SARS-CoV-2, as among the interferon-induced genes in nasal epithelial cells, with uncertain significance for SARS-CoV-2 infection[30, 71-73]. Indeed, we observe modest upregulation of this gene among cells from COVID-19 participants compared to healthy controls. Further, some of the cell subtypes identified as expanded during COVID-19 (e.g., interferon responsive ciliated cells, *BPIFA1*^{high} secretory cells, *BPIFA1*^{high}Chemokine^{high} secretory cells, and *KRT24*^{high}*KRT13*^{high} secretory cells) express relatively high abundances of *ACE2* (**Supplementary Figure 2.4E**).

Here, we discover that cells from individuals with mild/moderate COVID-19 recurrently upregulate interferon-responsive factors including *STAT1, MX1, HLA-B, HLA-C*, among others (compared to matched cell types among Control WHO 0 participants), while cells from individuals with severe COVID-19 repeatedly induced a distinct set of genes, including *S100A9, S100A8* and stress response factors (*HSPA8, HSPA1A, DUSP1,* **Figure 2.3H**).

We were curious as to whether depressed interferon and anti-viral responses could be explained by higher rates of steroid treatment among the severe COVID-19 group (**Table 2.1**). We therefore stratified our groups further into Steroid-Treated vs. Untreated, and assessed expression of genes previously identified as DE between Control WHO 0, COVID-19 WHO 1-5, and COVID-19 WHO 6-8. For some genes, steroid treatment partially suppressed the interferon response *within* each cohort – for instance, ciliated cells from Untreated COVID-19 WHO 1-5 participants showed higher abundances of *IFITM1, OAS2, IFI6*, and *IFI27* than their Steroid-Treated counterparts – while still maintaining strong differences in expression *between* groups (with abundance in COVID-19 WHO 1-5 > COVID-19 WHO 6-8 > COVID-19 WHO 0, see annotations on **Supplementary Figure 2.4C**). Interestingly, induction of *FKBP5* expression among ciliated cells from severe COVID-19 participants was fully explained by steroid treatment, which is consistent with the role for this protein in modulating glucocorticoid receptor activity. Other sets of anti-viral genes were equivalently expressed within each cohort, independent of steroid treatment, including *STAT1*, *STAT2, IFI44,* and *ISG15*. For many anti-viral factors in multiple cell types, we observed no effect of steroid treatment on the intrinsic anti-viral response during COVID-19.

Together, these data demonstrate global blunting of the anti-viral/interferon response among nasopharyngeal epithelial cells during severe COVID-19. We next attempted to query the source of local interferon, particularly in the COVID-19 WHO 1-5 samples where cell types appeared to be maximally responding to interferon stimulation. Notably, we expect many of the tissue-resident immune cells to reside principally within the deeper lamina propria and submucosal spaces, and are therefore are poorly represented in our dataset due to sampling type (swabbing of surface epithelial cells)[54]. Accordingly, we found exceedingly few immune cell types producing interferons: *IFNA* and *IFNB* were absent, rare *IFNL1* UMI were observed among T cells and Macrophages, and *IFNG* was robustly produced from Interferon Responsive Cytotoxic CD8 T cells, despite limited evidence for type II responses among epithelial cells (**Supplementary Figure 2.4F**). Further, we could not detect expression of any interferon types among epithelial cells, which is dramatically different from previous observations of robust type I/III interferon expression among nasal ciliated cells during influenza A and B infection[74] (**Supplementary Figure 2.4G**). Rather, we found robust induction of other inflammatory molecules from both immune and

epithelial cell types. *CXCL8* was produced by several specialized secretory cell types, including those uniquely expanded in COVID-19. Inflammatory macrophages and interferon responsive macrophages represent the primary sources of local *TNF*, *IL6*, and *IL10*, and uniquely express high abundances of chemoattractant molecules such as *CCL3, CCL2,* and *CXCL8.* Interestingly, interferon responsive macrophages appear to be a unique source of *CXCL9, CXCL10,* and *CXCL11* (**Supplementary Figures 2.4F**).

2.3.5 Targets of SARS-CoV-2 Infection in the Nasopharynx

Given a comprehensive picture of host cell biology during COVID-19 and across the spectrum of disease severity, we next tested whether the observed epithelial phenotypes were associated with altered local viral abundance. scRNA-seq protocols utilize poly-adenylated RNA capture and reverse transcription to generate snapshots of the transcriptional status of each individual cell. As other pathogens and commensal microbes also utilize poly-adenylation for RNA intermediates, or contain poly-adenylated stretches of RNA within their genomes, they may also be represented within scRNA-seq libraries. First, to perform an unbiased search for co-detected viral, bacterial, and fungal genomic material, we used metatranscriptomic classification to assign reads according to a comprehensive reference database[75,76] (previously described, see **Methods**). As expected, the majority (28/38) of swabs from individuals with COVID-19 contained reads classified as SARS coronavirus species (**Figure 2.4A, Supplementary Figures 2.5A-2.5C**). Among samples containing SARS coronavirus genomic material, the read abundance ranged from 2e0 to 8.8e6 reads (1.8e-3 to 1.9e4 reads/M total reads). We found little evidence for co-occurring respiratory viruses, which may be partially explained by the season when many of the swabs were collected (April-September 2020) and concurrent social distancing practices. Swabs from two individuals were found to contain rare reads classified as Influenza A virus species (maximum 5 reads per donor, within range for spurious classification), and we found no evidence for other seasonal human coronaviruses, Influenza B virus, metapneumovirus, or orthopneumovirus. Swabs from two individuals with mild/moderate COVID-19 were found to contain exceptionally high abundances of reads classified as Rhinovirus A (2.1e5 and 2.4e5 reads). Finally, we recovered low abundances of SARS coronavirus assigned reads from two participants from the Control WHO 0 cohort.

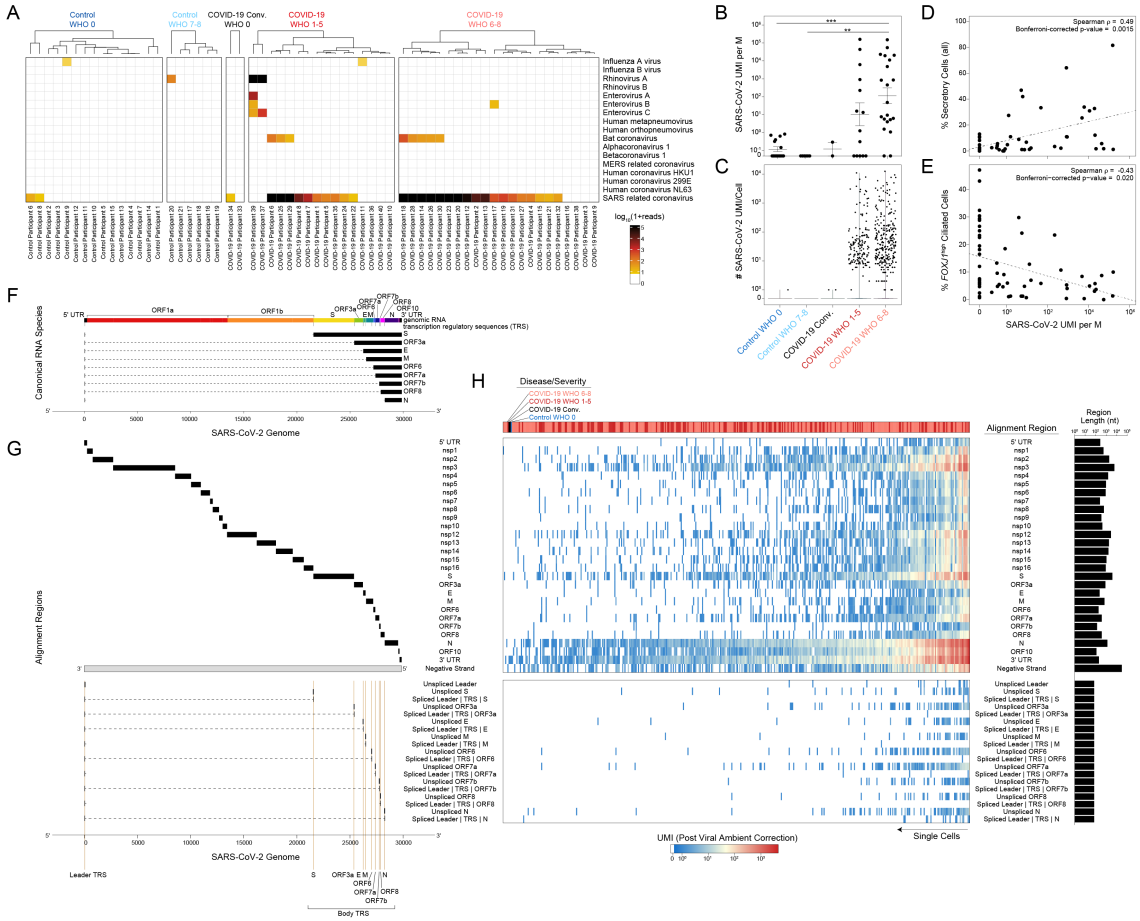


Figure 2.4: Co-detection of human and SARS-CoV-2 RNA
(A) Metatranscriptomic classification of all single-cell RNA-seq reads using Kraken2. Results shown from selected respiratory viruses. Only results with greater than 5 reads are shown. **(B)** Normalized abundance of SARS-CoV-2 aligning UMI from all single-cell RNA-seq reads (including those derived from ambient/low-quality cell barcodes). $P < 0.0001$ by Kruskal-Wallis test. Pairwise comparisons using Dunn's post-hoc testing. ** $p < 0.01$, *** $p < 0.001$. **(C)** Proportional abundance of Secretory cells (all) vs. total SARS-CoV-2 UMI (normalized to M total UMI). **(D)** Proportional abundance of *FOXJ1*^{high}Ciliated cells vs. total SARS-CoV-2 UMI (normalized to M total UMI). **(E)** SARS-CoV-2 UMI per high-quality cell barcode. Results following correction for ambient viral reads. **(F)** Schematic for SARS-CoV-2 genomic features annotated in the custom gtf. **(G)** Schematic for SARS-CoV-2 genome and subgenomic RNA species. **(H)** Heatmap of SARS-CoV-2 genes expression among SARS-CoV-2 RNA+ single cells (following correction for ambient viral reads). Top color bar indicates disease and severity cohort (red: COVID-19 WHO 1-5, pink: COVID-19 WHO 6-8, black: COVID-19 convalescent, blue: Control WHO 0). Top heatmap: SARS-CoV-2 genes and regions organized from 5' to 3'. Bottom heatmap: alignment to 70-mer regions directly surrounding viral transcription regulatory sequence (TRS) sites, suggestive of spliced RNA species (joining of the leader to body regions) vs. unspliced RNA species (alignment across TRS).

Next, we analyzed all SARS-CoV-2-aligned UMI following alignment to a joint genome containing both human and SARS-CoV-2[77]. We took the sum of all SARS-CoV-2 aligning UMI from a given participant – both from high-quality single-cell transcriptomes and low-quality/ambient RNA – as a representative measure of the total SARS-CoV-2 burden within the tissue microenvironment. As observed using metatranscriptomic classification, we found relatively low/spurious alignments to SARS-CoV-2 among Control participants, while swabs from

COVID-19 participants contained a wide range of SARS-CoV-2 aligning reads (**Figure 2.4B, Supplementary Figures 2.5D, 2.5E**). Samples from COVID-19 WHO 6-8 participants contained significantly higher abundances of SARS-CoV-2 aligning reads than both control cohorts, with an average of 1.1e2 +/- 2.8e0 (geometric mean +/- SEM) UMI per million aligned UMI (ranging from 0 to 1.5e5 per sample). Swabs from participants with mild/moderate COVID-19 contained slightly fewer SARS-CoV-2 aligning UMI, with an average of 1.1e1 +/- 4.3e0 (geometric mean +/- SEM) UMI per M.

Given the large diversity in SARS-CoV-2 abundance across all COVID-19 participants, we interrogated whether cell composition correlated with total SARS-CoV-2 (NB: contemporaneous work by our group has evaluated the accuracy of single-cell RNA-seq derived estimates of total SARS-CoV-2 abundance with more established protocols such as Real-Time RT-PCR). Among all cell types, we found that secretory cells were significantly positively correlated with the total viral abundance (Spearman's rho = 0.49, Bonferroni-corrected p = 0.0015), while *FOXJ1*^{high} ciliated cells were significantly negatively correlated (Spearman's rho = -0.43, Bonferroni-corrected p = 0.020, **Figures 2.4C, 2.4D**). This observation is in line with findings outlined in **Figures 1** and **2** where epithelial cell destruction during SARS-CoV-2 infection drives loss of mature ciliated cell types, which likely stimulates secretory cells to expand and repopulate lost epithelial cell types, although direct virally-mediated effects on secretory cell expansion have not been ruled out. Next, we binned the samples from COVID-19 participants into "Viral Low" and "Viral High" groupings (based on an arbitrary cutoff of 1e3 SARS-CoV-2 UMI per M, our findings were robust to a range of partition choices, **Supplementary Figures 2.5E, 2.5F**). Interferon responsive ciliated cells were expanded among "Viral High" COVID-19 samples and plasmacytoid DCs were absent from "Viral High" samples.

Next, we aimed to differentiate SARS-CoV-2 UMI derived from ambient or low-quality cell barcodes from those likely reflecting intracellular RNA molecules[74,78,79]. First, we filtered to only viral UMIs associated with cells presented in **Figure 2.1**, thereby removing those associated with low-quality or ambient-only cell barcodes (**Supplementary Figure 2.5G**). Using a combination of computational tools to 1) estimate the proportion of ambient RNA contamination per single cell and 2) estimate the abundance of SARS-CoV-2 RNA within the extracellular/ambient environment

(i.e., not cell-associated), we were able to test whether the amount of viral RNA associated with a given single-cell transcriptome was significantly higher than would be expected from ambient spillover. Together, this enabled us to identify cell barcodes whose SARS-CoV-2 aligning UMI were likely driven by spurious contamination, and annotate single cells that contain probable cell-associated or intracellular SARS-CoV-2 RNA (**Figure 2.4E, Supplementary Figure 2.5G**). Across all single cells, this analysis recovered 413 high-confidence SARS-CoV-2 RNA+ cells across 21 participants, and we confirmed that cell assignment as "SARS-CoV-2 RNA+" was not driven by technical factors such as sequencing depth or cell complexity (**Supplementary Figure 2.5H**). 262 cells were from participants with severe COVID-19 and 150 from mild/moderate COVID-19. We found one SARS-CoV-2 RNA+ cell from a participant with negative SARS-CoV-2 PCR. Among participants with any SARS-CoV-2 RNA+ cell, we found 20 +/- 7 (mean +/- SEM) SARS-CoV-2 RNA+ cells per sample (range 1-119), amounting to 4 +/-1.3% (range 0.1-24%) of the total recovered cells per sample. *Within* a given single cell, the abundance of SARS-CoV-2 UMI ranged from 1 to 12,612, corresponding to 0.01-98% of all human and viral UMI per cell.

To further understand the biological significance behind SARS-CoV-2 aligning UMI within a single cell, and to better identify cells with the highest-likelihood of actively supporting viral replication, we analyzed the specific viral sequences and their alignment regions in the viral genome[77,80,81]. During SARS-CoV-2 infection, viral uncoating from endosomal vesicles releases the positive, single-stranded, 5' capped, poly-adenylated genome into the host cytosol (**Figure 2.4F, 2.4G**). Here, translation of non-structural proteins proceeds first by templating directly off of the viral genome, generating a replication and transcription complex. The viral replication complex then produces both 1) negative strand genomic RNA intermediates, which serve as templates for further positive strand genomic RNA and 2) nested subgenomic mRNAs which are constructed from a 5' leader sequence fused to a 3' sequence encoding structural proteins for production of viral progeny (e.g., Spike, Envelope, Membrane, Nucleocapsid). Generation of nested subgenomic mRNAs relies on discontinuous transcription occurring between pairs of 6-mer transcriptional regulatory sequences (TRS), one 3' to the leader sequence (termed leader TRS, or TRS-L), and others 5' to each gene coding sequence (termed body TRS, or TRS-B)[82]. We reasoned that short SARS-CoV-2 aligning UMI could be readily distinguished by their strandedness (aligning to the negative vs. positive strand) and whether they fell within coding

regions, across intact TRS (indicating RNA splicing had not occurred for that RNA molecule at that splice site) or across a TRS with leader-to-body fusions (corresponding to subgenomic RNA, **Figure 2.4F, 2.4G, Supplementary Figure 2.6A**). Single cells containing higher abundances of spliced TRS or negative strand aligning reads are therefore more likely to represent truly virally-infected cells with a functional viral replication and transcription complex. Critically, the co-detection of host transcriptomic and viral genomic material associated with a single cell barcode cannot definitively establish the presence of intracellular virus and/or productive infection. Rather, below we integrate these and other aspects of the host and viral transcriptomes to refine and contextualize our confidence in "SARS-CoV-2 RNA+" cells.

The majority of SARS-CoV-2 aligning UMI among SARS-CoV-2 RNA+ cells was found heavily biased towards the 3' end of the genome, attributed to the 3' UTR, ORF10, and N gene regions, as expected due to poly-A priming (**Figure 2.4H**). A majority (68.7%) of SARS-CoV-2 RNA+ cells contained reads aligning to the viral negative strand, increasing the likelihood that many of these cells represent true targets of SARS-CoV-2 virions *in vivo*. In addition to negative strand alignment, we find roughly ~ 1/4 of the SARS-CoV-2 RNA+ cells contain at least 100 UMI that map to more than 20 distinct viral genomic locations per cell. Finally, comparing spliced to unspliced UMI, we found a minor fraction of cells with reads mapping directly across a spliced TRS sequence (4.6%), while 35% of SARS-CoV-2 RNA+ cells contained reads mapping across the equivalent 70mer window around an unspliced TRS. Notably, single cells containing reads aligning to spliced (subgenomic) RNA were heavily skewed toward those cells that contained the highest overall abundances of viral UMI – this may be an accurate reflection of coronavirus biology, wherein subgenomic RNA are most frequent within cells robustly producing new virions and total viral genomic material, but also points to inherent limitations in the detection of low-frequency RNA species by single-cell RNA-seq technologies.

Next, we integrated 1) the strand and splice information among SARS-CoV-2 aligning UMIs, 2) participant-to-participant diversity and 3) cell type annotations to gain a comprehensive picture of the identity and range of SARS-CoV-2 RNA+ cells within the nasopharyngeal mucosa (**Figure 2.5A-2.5D, Supplementary Figure 2.6A-2.6E**). We found incredible diversity in both the identity of SARS-CoV-2 RNA+ cells, as well as the distribution of SARS-CoV-2 RNA+ cells within and

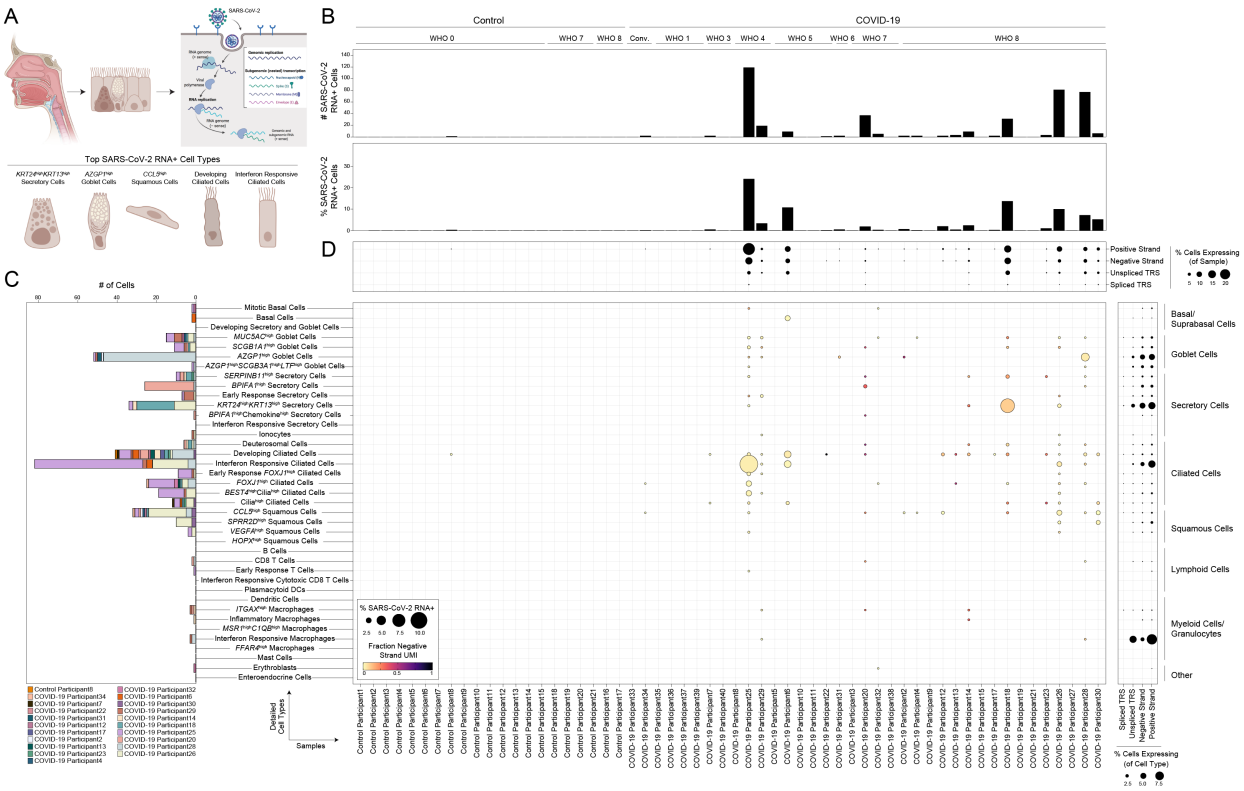


Figure 2.5: Cellular targets of SARS-CoV-2 in the nasopharynx
(A) Summary schematic of top SARS-CoV-2 RNA+ cells. (created with BioRender). **(B)** SARS-CoV-2 RNA+ cell abundance (top) and percent (bottom) per participant. Results following correction for ambient viral reads. **(C)** Abundance of SARS-CoV-2 RNA+ cells by detailed cell type, bars colored by participant. Results following correction for ambient viral reads. **(D)** Dot plot of SARS-CoV-2 RNA presence by sample (columns) and detailed cell types (rows). Dot size reflects fraction of a given participant and cell type containing SARS-CoV-2 RNA (following viral ambient correction). Dot color reflects fraction of aligned reads corresponding to the SARS-CoV-2 positive strand (yellow) vs. negative strand (black). Dot plot across columns: alignment of viral reads by participant, separated by RNA species type. Dot plot across rows: alignment of viral reads by detailed cell type, separated by RNA species type.

across participants. The majority of SARS-CoV-2 RNA+ cells were ciliated, goblet, secretory, or squamous. Highest-confidence SARS-CoV-2 RNA+ cells (spliced TRS UMI, negative strand UMI, > 100 SARS-CoV-2 UMI) tended to be found among *MUC5AC*^{high} goblet cells, *AZGP1*^{high} goblet cells, *BPIFA1*^{high} secretory cells, *KRT24*^{high}*KRT13*^{high} secretory cells, *CCL5*^{high} squamous cells, developing ciliated cells, and each ciliated cell subtype. A high proportion of interferon responsive macrophages contained SARS-CoV-2 genomic material, and rare *ITGAX*^{high} macrophages were found to contain UMI aligning to viral negative strand or spliced TRS regions – likely representing myeloid cells that have recently engulfed virally-infected epithelial cells or free virions. We did not find major differences in the presumptive cellular tropism by COVID-19 severity. A few cell types were commonly found to be SARS-CoV-2 RNA+ across all participants

(including participants with only rare viral RNA+ cells): most frequently, participants had at least one developing ciliated or squamous cell with SARS-CoV-2 RNA, followed by *MUC5AC*^{high} goblet cells, cilia^{high} ciliated cells, and *FOXJ1*^{high} ciliated cells (**Figure 2.5C**). However, among the individuals with the highest abundances of SARS-CoV-2 RNA+ cells, viral RNA was spread broadly across many different cell types, including those outside of the expected tropism for SARS-CoV-2 (e.g., also found within basal cells, ionocytes). Further, the cell types harboring the highest proportions of SARS-CoV-2 RNA+ cells represent the same cell types uniquely expanded or induced within COVID-19 participants, such as *KRT24*^{high}*KRT13*^{high} secretory cells, *AZGP1*^{high} goblet cells, and interferon responsive ciliated cells, and contain the highest abundances of *ACE2*-expressing cells (**Figure 2.5C, Supplementary Figure 2.6F**. Whether these cell types represent specific phenotypes elicited by intrinsic viral infection (potentially alongside induction of anti-viral genes) or are uniquely susceptible to SARS-CoV-2 entry (e.g., enhanced entry factor expression) will require further investigation. Finally, we compared the relative abundance of viral RNA *within* each cell type, and found developing ciliated cells contain significantly higher SARS-CoV-2 RNA molecules per-cell, including positive strand, negative strand-aligning reads, and spliced TRS reads (**Supplementary Figure 2.6G**). Intriguingly, among ciliated cell subtypes, interferon responsive ciliated cells, despite representing one of the most frequent "targets" of viral infection, contain the lowest per-cell abundances of SARS-CoV-2 RNA, potentially reflecting the impact of elevated anti-viral factors curbing high levels of intracellular viral replication (**Supplementary Figure 2.6H**).

2.3.6 Cell Intrinsic Responses to SARS-CoV-2 Infection

Above, we carefully mapped the specific cell types and states harboring SARS-CoV-2 RNA+ cells, identifying the subsets of epithelial cells that appear to actively support viral replication *in vivo* across distinct individuals (**Figure 2.5**). Further, we have characterized robust and cell-type-specific host responses among cells from COVID-19 participants, ostensibly representing both the bystander cell response to local virus and an inflammatory microenvironment, as well as the intrinsic response to intracellular SARS-CoV-2 RNA (**Figure 2.3**). Here, by directly comparing single cells containing SARS-CoV-2 RNA to their matched bystanders, we aimed to map both the cell-intrinsic response to direct viral infection, as well as the host cell identities that may *potentiate* or *enable* SARS-CoV-2 tropism and replication.

To control for variability among different SARS-CoV-2 RNA+ cell types and individuals, we compared SARS-CoV-2 RNA+ cells to bystander cells of the same cell type and participant. Among cell types with at least 5 SARS-CoV-2 RNA+ cells, we observed robust and specific transcriptional changes compared to both matched bystander cells as well as cells from healthy individuals (**Figures 2.6A, 2.6B)**. Notably, many of the genes previously identified as increased within all cells from COVID-19 participants, e.g., anti-viral factors *IFITM3, IFI44L*, were also upregulated among SARS-CoV-2 RNA+ cells compared to matched bystanders within multiple

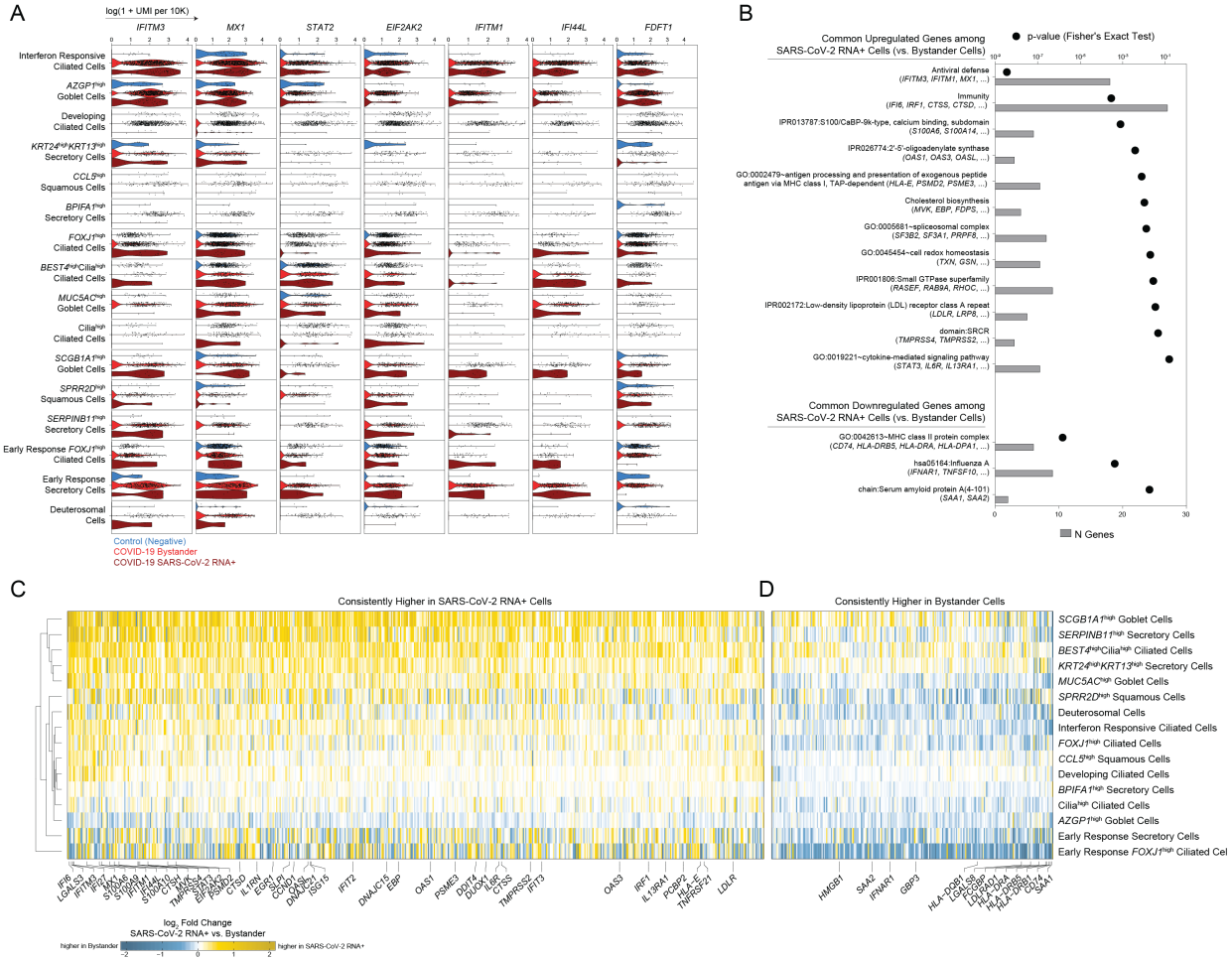


Figure 2.6: Intrinsic and bystander responses to SARS-CoV-2 infection

(A) Violin plot of selected genes upregulated in SARS-CoV-2 RNA+ cells in at least 3 individual cell type comparisons. Dark red: SARS-CoV-2 RNA+ cells, red: bystander cells from COVID-19 participants, blue: cells from Control participants. **(B)** Enriched gene ontologies among genes consistently up- or down-regulated among SARS-CoV-2 RNA+ cells across cell types. **(C)** Heatmap of genes consistently higher in SARS-CoV-2 RNA+ cells across multiple cell types. Colors represent log fold changes between SARS-CoV-2 RNA+ cells and bystander cells (SARS-CoV-2 RNA- cells, from COVID-19 infected donors) by cell type. Restricted to cell types with at least 5 SARS-CoV-2 RNA+ cells. Yellow: upregulated among SARS-CoV-2 RNA+ cells, blue: upregulated among bystander cells. **(D)** Heatmap of genes consistently higher in bystander cells across multiple cell types.

cell types. SARS-CoV-2 RNA+ cells from participants with mild/moderate COVID-19 showed stronger induction of anti-viral and interferon responsive pathways compared to those from participants with severe COVID-19, despite equivalent abundances of cell-associated viral UMI (**Supplementary Figure 2.7A**). *EIF2AK2,* which encodes protein kinase R and drives host cell apoptosis following recognition of intracellular double-stranded RNA, was among the most reliably expressed and upregulated genes among SARS-CoV-2 RNA+ cells compared to matched bystanders across diverse cell types, suggesting rapid activation of this locus following intrinsic PAMP recognition of SARS-CoV-2 replication intermediates[83]. Therefore, direct sensing of intracellular viral products amplifies interferon-responsive and anti-viral gene upregulation, though these pathways are also elevated within bystander cells.

The majority of genes induced within SARS-CoV-2 RNA+ cells were shared *across* diverse cell types, suggesting a conserved anti-viral response, as well as common features that facilitate or restrict infection (**Figure 2.6B-2.6D**). SARS-CoV-2 RNA appeared to robustly stimulate expression of genes involved in anti-viral sensing and defense (e.g., *MX1, IRF1, OAS1, OAS2)*, as well as genes involved in antigen presentation via MHC class I (**Figure 2.6C, Supplementary Table 2.5**). SARS-CoV-2 RNA+ cells expressed significantly higher abundances of multiple proteases involved in the cleavage of SARS-CoV-2 spike protein, a required step for viral entry (*TMPRSS4, TMPRSS2, CTSS, CTSD*). This suggests that within a given cell type, natural variations in the abundance of genes which support the viral life cycle partially account for which cells are successfully targeted by the virus. Among the core anti-viral/interferon-responsive gene sets induced within SARS-CoV-2 RNA+ cells, we found repeated and robust upregulation of *IFITM3* and *IFITM1*. Multiple studies have demonstrated that while these two interferon-inducible factors can disrupt viral release from endocytic compartments among a wide diversity of viral species, IFITMs can instead facilitate entry by human betacoronaviruses[80,84]. Therefore, enrichment of these factors within presumptive infected cells may reflect viral hijacking of a conserved host anti-viral responsive pathway. Genes involved in cholesterol and lipid biosynthesis were also upregulated among SARS-CoV-2 RNA+ cells, including *FDFT1*, *MVK, FDPS, ACAT2, HMGCS1*, all enzymes involved in the mevalonate synthesis pathway. In addition, SARS-CoV-2 RNA+ cells showed increased abundance of low-density lipoprotein receptors *LDLR* and *LRP8* compared to matched bystanders. Intriguingly, various genes involved in cholesterol metabolism

were recently identified as critical host factors for SARS-CoV-2 replication via CRISPR screens from multiple independent research groups[85,86]. Further, these groups found that direct inhibition of cholesterol biosynthesis decreased SARS-CoV-2 (as well as coronavirus strains 299E and OC43) replication within cell lines, and suggest S-mediated entry relies on host cholesterol. We queried the full collections of presumptive replication factors identified by four published CRISPR screens[85-88], and found significant enrichment among SARS-CoV-2 RNA+ cells for RAB GTPases (e.g. *RAB9A, RHOC, RASEF*), vacuolar ATPase H+ pump subunits, as well as transcriptional modulators such as *SPEN, SLTM, CREBBP, SMAD4 and EGR1* (**Supplementary Figure 2.7B**).

Finally, we found multiple previously-unappreciated genes implicated in susceptibility and response to SARS-CoV-2 infection, including S100/Calbindin genes such as *S100A6, S100A4,* and *S100A9*, which may directly play a role in leukocyte recruitment to infected cells. *IFNAR1* was substantially increased in many bystander cells compared to both cells from SARS-CoV-2 negative participants as well as matched SARS-CoV-2 RNA+ cells (**Figure 2.6D**). Blunting of interferon alpha signaling via downregulation of *IFNAR1* within SARS-CoV-2 RNA+ cells may partially explain high levels of viral replication compared to neighboring cells. Finally, bystander cells expressed significantly higher abundances of MHC-II molecules compared to SARS-CoV-2 RNA+ cells, including *HLA-DQB1, HLA-DRB1, HLA-DRB5, HLA-DRA,* and *CD74*.

2.4 Discussion

We have created a comprehensive map of SARS-CoV-2 infection of the human nasopharynx using scRNA-seq. We hypothesize that the host response at the site of initial infection, the nasal mucosa, is an essential determinant of overall COVID-19 disease trajectory. By dissecting the nature of host-pathogen interactions at this primary viral target across the spectrum of disease outcomes, we can characterize both protective and pathogenic responses to SARS-CoV-2 infection. Here, we begin to untangle the myriad factors that may restrict viral infection to the upper respiratory tract or support the development of severe lower respiratory tract disease.

First, we find that mature ciliated cells decline dramatically within the nasopharynx of COVID-19 samples, directly correlated with the tissue abundance of SARS-CoV-2 RNA at the time of sampling. Conversely, secretory cell populations expand among samples with high viral loads,

which potentially represents a conserved response for epithelial re-population of lost mature ciliated cells through a recently-identified mechanism of secretory/goblet differentiation, using deuterosomal cells as intermediates[54,55]. Accordingly, deuterosomal cells and immature/developing ciliated cells were considerably expanded among COVID-19 samples, suggesting interdependence between each of these compartments in maintaining epithelial homeostasis during viral challenge. SARS-CoV-2 infection also induced dramatic increases in the diversity of epithelial cell types, both with respect to shifted compositional balance among major cell identities, and also via expansion of specialized secretory and goblet cell subsets, including a subset termed *KRT24*^{high}*KRT13*^{high} secretory cells, which closely match the recently-identified KRT13+ "hillock" cell, previously associated with epithelial regions experiencing rapid cellular turnover and inflammation[54,57,58]. Other specialized subsets of secretory and goblet cells, such as early response secretory cells, *AZGP1*^{high} goblet cells, and *SCGB1A1*^{high} goblet cells, are expanded among COVID-19 participants. However, expansion of these cells is observed within discrete subsets of individuals and is not homogenous across the disease groups we sampled here. Indeed, understanding whether heterogeneous responses in the epithelial compartment *between* individuals with COVID-19 underscores differences in disease manifestations will require larger cohort studies, with a focus on longitudinal responses following initial infection. Indeed, further work is required to understand how the epithelial responses to SARS-CoV-2 infection within the nasal mucosa relates to epithelial responses in other common upper respiratory viral infections and inflammatory states.

Beyond cellular compositional changes during COVID-19, our study identified marked variability in the induction of anti-viral gene expression that was associated with disease severity. We found robust upregulation of interferon stimulated genes among epithelial and immune cells isolated from individuals with mild/moderate COVID-19, and this was particularly evident in cells that contained SARS-CoV-2 RNA. Surprisingly, despite strong induction of anti-viral gene expression, we found little to no mRNA corresponding to type I or type III interferons amongst any recovered cell types. In a related study mapping the nasal epithelium during influenza infection, we and our colleagues found extensive upregulation of *IFNA*, *IFNB1,* and *IFNL1-3* within ciliated cells and goblet cells, both highlighting the capacity of superficial nasal epithelial cells to secrete local interferons during viral infection, but also the technical capacity of the scRNA-seq platform used

in both studies to capture interferon mRNA[74]. The precise sources and signals which motivate a broad anti-viral response among mild COVID-19 cases in our study remains unknown – they may originate from immune cells contained deeper within the respiratory mucosa (therefore inaccessible through the superficial sampling used here), from sparse, highly transient interferon expression from superficial epithelial or immune cells, or may derive from direct PAMP/DAMP sensing or alternative inflammatory signals.

Remarkably, in comparison to individuals with mild/moderate disease, we found that anti-viral gene expression was profoundly blunted in cells isolated from individuals with severe disease, even in cells containing SARS-CoV-2 RNA. This effect was observed among diverse cell types, including those thought to represent direct targets of viral infection, such as ciliated cells and secretory cells, and also bystanders and co-resident immune cells. Notably, individuals with severe COVID-19 disease had equivalent or even elevated levels of nasal SARS-CoV-2 RNA at the time of sampling, and contained expanded inflammatory and type II-interferon responsive macrophages compared to mild/moderate cases. Indeed, published peripheral immune studies comparing mild and severe COVID-19 also observe diminished type I and type III interferon abundances in severe cases, and note restricted interferon stimulated gene expression among circulating immune cells[17,18,22]. Other human betacoronaviruses including MERS and SARS-CoV exhibit multiple strategies to avoid triggering pattern recognition receptor pathways, including degradation of host mRNA within infected cells[89,90], sequestration of viral replication intermediates (e.g., double stranded RNA) from host sensors[91], and direct inhibition of immune effector molecules[80,83,92], thereby leading to diminished induction of anti-viral pathways and blunted autocrine and paracrine interferon signaling. Strategies to avoid innate immune recognition have now been extended to SARS-CoV-2 as well, indicating that avoiding host recognition is likely an essential aspect of viral success[93-95]. The close association we observe between disease severity and weak anti-viral gene expression among nasal epithelial cells is intriguing given recent observations of inborn defects in TLR3, IRF7, IRF9, and IFNAR1, or antibody-mediated neutralization of type I interferon responses within individuals who develop severe COVID-19[49-51]. Further, we found that lower nasal viral loads were associated with elevated detection of tissue plasmacytoid DCs, suggesting diminished or delayed recruitment of these cells may partially explain how local viral replication proceeds to such high abundances. Taken together, these findings strongly suggest that severe

infection can arise in the setting of an intrinsic impairment of epithelial anti-viral immunity, and that timely induction of anti-viral responses are an essential aspect of successful viral control. We surmise that the combined effects of a viral strain with naturally poor interferon induction and intrinsic defects in immune or epithelial anti-viral responses within the nasal mucosa may predispose to severe disease via prolonged viral replication in the upper airway, eventually leading to immunopathology characteristic of severe COVID-19.

Critically, our work does not address the dynamics of nasal epithelial anti-viral responses during SARS-CoV-2 infection in individual patients, nor does it directly relate intrinsic mucosal responses in the nasopharynx to potential interferon or anti-viral responses in the lung or distal airways. Indeed, related work suggests type III interferons are present in the lungs, but not the nasopharynx, during SARS-CoV-2 infection, and may contribute to tissue damage late in disease course[96]. Further, as the individuals in our cohort were intentionally sampled as early within their disease course as possible and the majority have elevated viral levels within their nasopharynx, our findings have an unclear relation to the tissue response during hyper-inflammatory "late" stages of COVID-19. However, among individuals who develop severe COVID-19, we observe unique recruitment of highly inflammatory macrophages that represent the major tissue sources of proinflammatory cytokines including *IL1B*, *TNF*, *CXCL8*, *CCL2*, *CCL3* and *CXCL9/10/11* – of likely relation to the immune dysregulation characterized by elevation of the same factors in the periphery in late, severe disease. In addition, we note specific upregulation of alarmins *S100A8/S100A9* (i.e., calprotectin) among epithelial cells in severe COVID-19 compared to mild and control counterparts, and even higher expression of *S100A9* within SARS-CoV-2 RNA+ cells from those same individuals. A recent study identified these as potential biomarkers of severe COVID-19, and proposed that these factors directly drive excessive inflammation and precede the massive cytokine release characteristic of late disease[97]. Our work suggests that severe COVID-19-specific expression of calprotectin may originate instead within the virally-infected nasal epithelia, and suggests that further work to understand the epithelial cell regulation of *S100A8/A9* gene expression may help clarify maladaptive responses to SARS-CoV-2 infection.

Finally, we provide a direct investigation into the host factors that enable or restrict SARS-CoV-2 replication within epithelial cells *in vivo*. Here, we recapitulate expected "hits" based on well-

described host factors involved in viral replication – e.g., *TMPRSS2*, *TMPRSS4* enrichment among presumptive virally infected cells. We similarly observed expression of anti-viral genes which were globally enriched among cells from mild/moderate COVID-19 participants, with even higher expression among the viral RNA+ cells themselves. In accordance with previous studies into the nasal epithelial response to influenza infection[74], we observed bystander epithelial cell upregulation of both MHC-I and MHC-II family genes; however, we found that SARS-CoV-2 RNA+ cells only expressed MHC-I, and uniformly downregulated MHC-II genes compared to matched bystanders. To our knowledge, downregulation of host cell pathways for antigen presentation by coronaviruses has not been previously described. A recent study found that CIITA and CD74 can intrinsically block entry of a range of viruses (including SARS-CoV-2) via endosomal sequestration, and therefore cells that upregulate these (and other) components of MHC-II machinery may naturally restrict viral entry[98].

Together, our work demonstrates that many of the factors associated with the clinical trajectory following SARS-CoV-2 infection stem from initial host-viral encounters in the nasopharyngeal epithelium. Further, it suggests that there may be a clinical window in which severe disease can be subverted by focusing preventative or therapeutic interventions early within the nasopharynx, thereby bolstering anti-viral responses and curbing pathological inflammatory signaling prior to development of severe respiratory dysfunction or systemic disease.

2.5 Methods

2.5.1 Study Participants and Design

Eligible participants were recruited from to the University of Mississippi Medical Center (UMMC) outpatient clinics, medical surgical units, Intensive Care Units (ICU), or endoscopy units between April 2020 and September 2020. The UMMC Institutional Review Board approved the study under IRB#2020-0065. All participants, or their legally authorized representative provided written informed consent. Participants were eligible for inclusion in the COVID-19 cohort if they were at least 18 years old, had a positive nasopharyngeal swab for SARS-CoV-2 by PCR, had COVID-19 related symptoms including fever, chills, cough, shortness of breath, and sore throat, and weighed more than 110 lb. Participants were eligible for the Control cohort if they were at least 18 years old, had a current negative SARS-CoV-2 test (PCR or rapid antigen test), and weighed more than

110 lb. Exclusion criteria for both cohorts included a history of blood transfusion within 4 weeks and subjects who could not be assigned a definitive COVID-19 diagnosis from nucleic acid testing. 38 individuals with COVID-19 were included, both male (n = 20) and female (n = 18). For the Control cohort, 21 participants were included – 11 identified as male, 10 as female. The median age of COVID-19 participants was 56.5 years old; the median age of Control participants was 62 years old. Among hospitalized participants, samples were collected between Day 1 to Day 3 of hospitalization. COVID-19 participants were classified according to the 8-level ordinal scale proposed by the WHO representing their peak clinical severity and level of respiratory support required.

2.5.2 Sample Collection and Biobanking

Nasopharyngeal samples were collected by trained healthcare provider using FLOQSwabs (Copan flocked swabs) following the manufacturer's instructions. Collectors would don personal protective equipment (PPE), including a gown, non-sterile gloves, a protective N95 mask, a bouffant, and a face shield. The patient's head was then tilted back slightly, and the swab inserted along the nasal septum, above the floor of the nasal passage to the nasopharynx until slight resistance was felt. The swab was then left in place for several seconds to absorb secretions and slowly removed while rotating swab. The swabs were then placed into a cryogenic vial with 900 µL of heat inactivated fetal bovine serum (FBS) and 100 µL of dimethyl sulfoxide (DMSO). The vials were then placed into a Thermo Scientific Mr. Frosty Freezing Container for optimal cell preservation. The Mr. Frosty containing the vials was then placed in cooler with dry ice for transportation from patient area to laboratory for processing. Once in the laboratory, the Mr. Frosty was placed into the -80°C freezer overnight and then on the next day, the vials were moved to the liquid nitrogen storage container.

2.5.3 Dissociation and Collection of Viable Single cells from Nasal Swabs

Swabs in freezing media (90% FBS/10% DMSO) were stored in liquid nitrogen until immediately prior to dissociation. A detailed sample protocol can be found here: https://protocols.io/view/human-nasopharyngeal-swab-processing-for-viable-si-bjhkkj4w.html.[99]. This approach ensures that all cells and cellular material from the nasal swab (whether directly attached to the nasal swab, or released during the washing and digestion process), are exposed first

to DTT for 15 minutes, followed by an Accutase digestion for 30 minutes. Briefly, nasal swabs in freezing media were thawed, and each swab was rinsed in RPMI before incubation in 1 mL RPMI/10 mM DTT (Sigma) for 15 minutes at 37ºC with agitation. Next, the nasal swab was incubated in 1 mL Accutase (Sigma) for 30 minutes at 37ºC with agitation. The 1 mL RPMI/10 mM DTT from the nasal swab incubation was centrifuged at 400 g for 5 minutes at 4ºC to pellet cells, the supernatant was discarded, and the cell pellet was resuspended in 1 mL Accutase and incubated for 30 minutes at 37ºC with agitation. The original cryovial containing the freezing media and the original swab washings were combined and centrifuged at 400 g for 5 minutes at 4ºC. The cell pellet was then resuspended in RPMI/10 mM DTT, and incubated for 15 minutes at 37ºC with agitation, centrifuged as above, the supernatant was aspirated, and the cell pellet was resuspended in 1 mL Accutase, and incubated for 30 minutes at 37ºC with agitation. All cells were combined following Accutase digestion and filtered using a 70 µm nylon strainer. The filter and swab were washed with RPMI/10% FBS/4 mM EDTA, and all washings combined. Dissociated, filtered cells were centrifuged at 400 g for 10 minutes at 4ºC, and resuspended in 200 µL RPMI/10% FBS for counting. cells were diluted to 20,000 cells in 200 µL for scRNA-seq. For the majority of swabs, fewer than 20,000 cells total were recovered. In these instances, all cells were input into scRNA-seq.

2.5.4 scRNA-seq

Seq-Well S^3 was run as previously described[52,53,100]. Briefly, a maximum of 20,000 single cells were deposited onto Seq-Well arrays preloaded with a single barcoded mRNA capture bead per well. cells were allowed to settle by gravity into wells for 10 minutes, after which the arrays were washed with PBS and RPMI, and sealed with a semi-permeable membrane for 30 minutes, and incubated in lysis buffer (5 M guanidinium thiocyanate/1 mM EDTA/1% BME/0.5% sarkosyl) for 20 minutes. Arrays were then incubated in a hybridization buffer (2M NaCl/8% v/v PEG8000) for 40 minutes, and then the beads were removed from the arrays and collected in 1.5 mL tubes in wash buffer (2M NaCl/3 mM $MgCl_2$/20 mM Tris-HCl/8% v/v PEG8000). Beads were resuspended in a reverse transcription master mix, and reverse transcription, exonuclease digestion, second strand synthesis, and whole transcriptome amplification were carried out as previously described. Libraries were generated using Illumina Nextera XT Library Prep Kits and sequenced on NextSeq

500/550 High Output v2.5 kits to an average depth of 180 million aligned reads per array: read 1: 21 (cell barcode, UMI), read 2: 50 (digital gene expression), index 1: 8 (N700 barcode).

2.5.5 Data Preprocessing and Quality Control

Pooled libraries were demultiplexed using bcl2fastq (v2.17.1.14) with default settings (mask_short_adapter_reads 10, minimum_trimmed_read_length 10, implemented using Cumulus, snapshot 4, https://cumulus.readthedocs.io/en/stable/bcl2fastq.html)[101]. Libraries were aligned using STAR within the Drop-Seq Computational Protocol (https://github.com/broadinstitute/Drop-seq) and implemented on Cumulus (https://cumulus.readthedocs.io/en/latest/drop_seq.html, snapshot 9, default parameters)[102]. A custom reference was created by combining human GRCh38 (from cellRanger version 3.0.0, Ensembl 93) and SARS-CoV-2 RNA genomes. The SARS-CoV-2 viral sequence and GTF are as described in Kim et al. 2020 (https://github.com/hyeshik/sars-cov-2-transcriptome, BetaCov/South Korea/KCDC03/2020 based on NC_045512.2)[77]. The GTF includes all CDS regions (as of this annotation of the transcriptome, the CDS regions completely cover the RNA genome without overlapping segments), and regions were added to describe the 5' UTR ("SARSCoV2_5prime"), the 3' UTR ("SARSCoV2_3prime"), and reads aligning to anywhere within the Negative Strand ("SARSCoV2_NegStrand"). Trailing A's at the 3' end of the virus were excluded from the SARS-CoV-2 FASTA, as these were found to drive spurious viral alignment in pre-COVID19 samples. Finally, additional small sequences were appended to the FASTA and GTF that differentiate reads that align to the 70-nucleotide region around the viral TRS sequence – either across the intact, unspliced genomic sequences (e.g. named "SARSCoV2_Unspliced_S" or "SARSCoV2_Unspliced_Leader") or various spliced RNA species (e.g. "SARSCoV2_Spliced_Leader_TRS_S"), see schematics in **Figures 2.4F, 2.4G, Supplementary Figures 2.6A**. Alignment references were tested against a diverse set of pre-COVID-19 samples and *in vitro* SARS-CoV-2 infected human bronchial epithelial cultures[36] to confirm specificity of viral aligning reads. Aligned cell-by-gene matrices were merged across all study participants, and cells were filtered to eliminate barcodes with fewer than 200 UMI, 150 unique genes, and greater than 50% mitochondrial reads. Of the 59 nasal swabs thawed and processed, 3 contained no high-quality cell barcodes after sequencing (NB: these samples contained < 5,000 viable cells prior to Seq-Well array loading).This resulted in a final dataset of

32,871 genes and 32,588 cells across 56 study participants (35 COVID-19 individuals, 21 control individuals). Preprocessing, alignment, and data filtering was applied equivalently to samples from the fresh vs. frozen cohort. For analysis of RNA velocity, we also recovered both exonic and intronic alignment information using DropEst (Cumulus (https://cumulus.readthedocs.io/en/latest/drop_seq.html, snapshot 9, dropest_velocyto true, run_dropest true)[103].

2.5.6 Cell Clustering and Annotation

Dimensionality reduction, cell clustering, and differential gene analysis were all achieved using the Seurat (v3.1.5) package in R programming language (v3.0.2)[104]. Dimensionality reduction was carried out by running principal components analysis over the 3,483 most variable genes with dispersion > 0.8 (tested over a range of dispersion > 0.7 to dispersion > 1.2; dispersion > 0.8 was determined as optimal based on number of variable genes, and general stability of clustering results across these cutoffs was confirmed). Only variable genes from human transcripts were considered for dimensionality reduction and clustering. Using the Jackstraw function within Seurat, we selected the first 36 principal components that described the majority of variance within the dataset, and used these for defining a nearest neighbor graph and Uniform Manifold Approximation and Projection (UMAP) plot. Cells were clustered using Louvain clustering, and the resolution parameter was chosen by maximizing the average silhouette score across all clusters. Differentially expressed genes between each cluster and all other cells were calculated using a likelihood ratio test, implemented with Seurat's FindAllMarkers function, test.use set to "bimod"[105]. Clusters were merged if they failed to contain sets of significantly differentially expressed genes. We proceeded iteratively through each cluster and subcluster until "terminal" cell subsets/cell states were identified – we defined "terminal" cell states when principal components analysis and Louvain clustering did not confidently identify additional sub-states, as measured by abundance of differentially expressed genes between potential clusters. For visualization in **Figure 2.2**, we pooled all cells determined to be of epithelial origin, and using the methods for dimensionality reduction as above (dispersion cutoff > 1, 30 principal components). We applied similar approaches for immune cell types (**Supplementary Figure 2.3**), including iterative subclustering to resolve and annotate all constituent cells types and subtypes. Gene module scores were calculated using the AddModuleScore function within Seurat.

We annotated epithelial subtypes according to the following groups and representative markers: goblet cells were split into 4 distinct sets: *MUC5AC*^{high} goblet cells, which lacked additional specialized markers beyond classic goblet cell identifiers, *SCGB1A1*^{high} goblet cells, *AZGP1*^{high} goblet cells, and *AZGP1*^{high}*SCGB3A1* ^{high}*LTF*^{high} goblet cells. Secretory cells were divided into 6 distinct detailed subtypes: *SERPINB11*^{high} secretory cells (which, similar to *MUC5AC*^{high} goblet cells, represented a more "generic" or un-differentiated secretory cell phenotype), *BPIFA1*^{high} secretory cells, early response secretory cells (which expressed genes such as *JUN, EGR1, FOS, NR4A1*), *KRT24*^{high}*KRT13*^{high} secretory cells, *BPIFA1*^{high}Chemokine^{high} secretory cells (chemokines include *CXCL8, CXCL2, CXCL1,* and *CXCL3),* and interferon responsive secretory cells (defined by higher expression of broad anti-viral genes including *IFITM3, IFI6,* and *MX1*). Subsets of squamous cells were also found – detailed squamous cell subtypes include *CCL5*^{high} squamous cells, *VEGFA*^{high} squamous cells (which express multiple vascular endothelial genes including *VEGFA and VWF*), *SPRR2D*^{high} squamous cells (which, in addition to *SPRR2D*, express the highest abundances of multiple SPRR- genes including *SPRR2A, SPRR1B, SPRR2E,* and *SPRR3*), and *HOPX*^{high} squamous cells. Finally, ciliated cells could be further divided into 5 distinct subtypes: interferon responsive ciliated cells (expressing anti-viral genes similar to other "interferon responsive" subsets, such as *IFIT1, IFIT3, IFI6*), *FOXJ1*^{high} ciliated cells, early response *FOXJ1*^{high} ciliated cells (which, in addition to high *FOXJ1*, also express higher abundances of genes such as *JUN*, *EGR1*, *FOS* than other ciliated cell subtypes), cilia^{high} ciliated cells (which broadly express the highest abundances of structural cilia genes, such as *DLEC1* and *CFAP100),* and *BEST4*^{high}cilia^{high} ciliated cells (in addition to cilia components, also express the ion channel *BEST4*).

2.5.7 RNA Velocity and Pseudotemporal Ordering of Epithelial cells

RNA velocity was modeled using the scVelo package, version 0.2.3[68,69]. Using cluster annotations previously assigned from iterative clustering in Seurat, cells from epithelial cell types were pre-processed according to the scVelo pipeline: genes were normalized using default parameters (pp.filter_and_normalize), principal components and nearest neighbors in PCA space were calculated (using defaults of 30 PCs, 30 nearest neighbors), and the first and second order moments of nearest neighbors were computed, which are used as inputs into velocity estimates

(pp.moments). RNA velocity was estimated using the scVelo tool tl.recover_dynamics with default input parameters, which maps the full splicing kinetics for all genes and tl.velocity, with mode='dynamical'. Top velocity transition "driver" genes were identified by high "fit_likelihood" parameters from the dynamical model, and are used for visualization in **Supplementary Figure 2.2C**. The same approaches were used for modeling RNA velocity among only Basal, secretory, and goblet cells (**Figures 2.2F-2.2I**), only ciliated cells (**Figures 2.2J-2.2M**), and only COVID-19 or only Control cells (**Figures 2.2P, 2.2Q**). For RNA velocity analysis of ciliated cells or Basal, secretory and goblet cells, the velocity pseudotime was calculated using the tl.velocity_pseudotime function with default settings.

2.5.8 Metatranscriptomic Classification of Reads from Single-cell RNA-Seq

To identify co-detected microbial taxa present in the cell-associated or ambient RNA of nasopharyngeal swabs, we used the Kraken2 software implemented using the Broad Institute viral-ngs pipelines on Terra (https://github.com/broadinstitute/viral-pipelines/tree/master)[76]. A previously-published reference database included human, archaea, bacteria, plasmid, viral, fungi, and protozoa species and was constructed on May 5, 2020, therefore included sequences belonging to the novel SARS-CoV-2 virus[75]. Inputs to Kraken2 were: kraken2_db_tgz = "gs://pathogen-public-dbs/v1/kraken2-broad-20200505.tar.zst", krona_taxonomy_db_kraken2_tgz = "gs://pathogen-public-dbs/v1/krona.taxonomy-20200505.tab.zst", ncbi_taxdump_tgz = "gs://pathogen-public-dbs/v1/taxdump-20200505.tar.gz", trim_clip_db = "gs://pathogen-public-dbs/v0/contaminants.clip_db.fasta" and spikein_db = "gs://pathogen-public-dbs/v0/ERCC_96_nopolyA.fasta". Species with fewer than 5 reads were considered spurious and excluded.

2.5.9 Correction for Ambient Viral RNA

Single-cell data from high-throughput single-cell RNA-seq platforms frequently experience low-levels of non-specific RNA assigned to cell barcodes that does not represent true cell-derived transcriptomic material, but rather contamination from the ambient pool of RNA. To safeguard against spurious assignment of SARS-CoV-2 RNA to cells without true intracellular viral material, i.e., viral RNA non-specifically picked up from the microenvironment as a component of ambient RNA contamination, we employed the following corrections and statistical tests to control for

ambient viral RNA and enable confident assignments for SARS-CoV-2 RNA+ cells. Similar to approaches previously described, we tested whether the abundance of viral RNA within a given single cell was significantly higher than expected by chance given the estimate of ambient RNA contaminating that cell, as well as the proportion of viral RNA of the total ambient RNA pool[74,78,79]. First, this required modeling and estimating the ambient RNA fraction associated with each individual swab. Here, we employed CellBender (https://github.com/broadinstitute/CellBender), a software package built to learn the ambient RNA profile and provide an ambient RNA-corrected output[78]. Input UMI count matrices contained the top 10,000 cell barcodes, therefore including at least 70% cell barcodes sampling the ambient RNA of low-quality cell pool. CellBender's remove-background function was run with default parameters and --fpr 0.01 --expected-cells 500 --low-count-threshold 5. Using the corrected output from each sample's count matrix following CellBender, we calculated the proportion of ambient contamination per high-quality cell by comparing to the single-cell's transcriptome pre-correction, and summed all UMI from background/low-quality cell barcodes to recover an estimate of the total ambient pool. Next, we tested whether the abundance of viral RNA in a given single cell was significantly above the null abundance given the ambient RNA characteristics using an exact binomial test (implemented in R (binom.test):

$P(x) = \frac{n!}{(n-x)!x!} p^x q^{n-x}$ where n = SARS-CoV-2 UMI per cell, x = total UMI per cell

p = (ambient fraction per cell)*(SARS-CoV-2 UMI fraction of all ambient UMI), and q = 1-p

P-values were FDR-corrected within sample, and cells whose SARS-CoV-2 UMI abundance with FDR < 0.01 were considered "SARS-CoV-2 RNA+".

2.5.10 Differential Expression by Cohort, Cell Type, or Viral RNA Status

To compare gene expression between cells from distinct donor cohorts we employed a negative binomial generalized linear model. Cells from each cell type belonging to either COVID-19 WHO 1-5 (mild/moderate), COVID-19 WHO 6-8 (severe), or Control WHO 0 were compared in a pairwise manner, implemented using the Seurat FindAllMarkers function (test.use = "negbinom"). We considered genes as differentially expressed with an FDR-adjusted p value < 0.001 and log fold change > 0.25. To compare gene expression between SARS-CoV-2 RNA+ cells and bystander

cells (from COVID-19 participants, but without intracellular viral RNA) we again used a negative binomial generalized linear model, but instead implemented using DESeq2[106]. We only tested cell types containing at least 15 SARS-CoV-2 RNA+ cells, and for each cell type, we restricted our bystander cells to the same participants as the SARS-CoV-2 RNA+ cells. Next, given the large discrepancies in cell number between SARS-CoV-2 RNA+ and bystander groups among most cell types, we randomly sub-sampled the bystander cells to at most 4x the number of SARS-CoV-2 RNA+ cells. Further, we selected bystander cell subsets that matched the cell quality distribution of the SARS-CoV-2 RNA+ cells, based on binned deciles of UMI/cell. DESeq2 was run with default parameters and test = "Wald". Gene ontology analysis was run using the Database for Annotation, Visualization, and Integrated Discovery (DAVID)[107]. Gene set enrichment analysis (GSEA) was completed using the R package fgsea over genes ranked by average log foldchange expression between each cohort, including all genes with an average expression > 0.5 UMI within each respective cell type[108]. Gene lists corresponding to "Shared IFN Response", "Type I IFN Specific Response" and "Type II IFN Specific Response" are derived from previously-published population RNA-seq data from nasal epithelial basal cells treated *in vitro* with 0.1 ng/mL – 10 ng/mL IFNα or IFNγ for 12 hours[30]. Module scores were calculated using the Seurat function AddModuleScore with default inputs.

2.5.11 Statistical Testing

All statistical tests were implemented either in R (v4.0.2) or Prism (v6) software[109]. Comparisons between cell type proportions by cohort were tested using a Kruskal-Wallis test with bonferroni-correction, implemented in R using the kruskal.test, and p.adjust functions. Post-tests for between-group pairwise comparisons used Dunn's test. Spearman correlation was used where appropriate, implemented using the cor.test function in R. All testing for differential expression was implemented in R using either Seurat, scVelo, or DESeq2, and all results were FDR-corrected as noted in specific **Methods** sections. P-values, n, and all summary statistics are provided either in the results section, figure legends, figure panels, or supplementary tables.

2.5.12 Data and Code Availability

Prism (v6), R (v4.0.2) packages ggplot2 (v3.3.2[110]), Seurat (v3.2.2[111]), ComplexHeatmap (v2.7.3[112]), and Circlize (0.4.11[113]), fgsea (v.1.16.0[108]) and Python (v3.8.3) package scVelo

(v0.3.0[68]) were used for visualization. All raw, normalized, and annotated data is available for download and visualization via the Single cell Portal: https://singlecell.broadinstitute.org/single_cell/study/SCP1289/impaired-local-intrinsic-immunity-to-sars-cov-2-infection-underlies-severe-covid-19. Interim data was also deposited in a single-cell data resource for COVID-19 studies: https://www.covid19cellatlas.org[114]. Custom reference FASTA and GTF for SARS-CoV-2 is available for download:

2.6: References

1. Pan, Y., Zhang, D., Yang, P., Poon, L.L.M., and Wang, Q. (2020). Viral load of SARS-CoV-2 in clinical samples. Lancet Infect. Dis.
2. Sanche, S., Lin, Y.T., Xu, C., Romero-Severson, E., Hengartner, N., and Ke, R. (2020). RESEARCH High Contagiousness and Rapid Spread of Severe Acute Respiratory Syndrome Coronavirus 2. Emerg. Infect. Dis.
3. Fears, A.C., Klimstra, W.B., Duprex, P., Hartman, A., Weaver, S.C., Plante, K.S., Mirchandani, D., Plante, J.A., Aguilar, P. V., Fernández, D., et al. (2020). Persistence of Severe Acute Respiratory Syndrome Coronavirus 2 in Aerosol Suspensions. Emerg. Infect. Dis.
4. Meyerowitz, E.A., Richterman, A., Gandhi, R.T., and Sax, P.E. (2021). Transmission of SARS-CoV-2: A Review of Viral, Host, and Environmental Factors. Ann. Intern. Med.
5. Arons, M.M., Hatfield, K.M., Reddy, S.C., Kimball, A., James, A., Jacobs, J.R., Taylor, J., Spicer, K., Bardossy, A.C., Oakley, L.P., et al. (2020). Presymptomatic SARS-CoV-2 Infections and Transmission in a Skilled Nursing Facility. N. Engl. J. Med.
6. Sakurai, A., Sasaki, T., Kato, S., Hayashi, M., Tsuzuki, S., Ishihara, T., Iwata, M., Morise, Z., and Doi, Y. (2020). Natural History of Asymptomatic SARS-CoV-2 Infection. N. Engl. J. Med.
7. Wang, Y., Liu, Y., Liu, L., Wang, X., Luo, N., and Ling, L. (2020b). Clinical outcome of 55 asymptomatic cases at the time of hospital admission infected with SARS-Coronavirus-2 in Shenzhen, China. J. Infect. Dis.
8. Guan, W., Ni, Z., Hu, Y., Liang, W., Ou, C., He, J., Liu, L., Shan, H., Lei, C., Hui, D.S.C., et al. (2020). Clinical Characteristics of Coronavirus Disease 2019 in China. N. Engl. J. Med.
9. Huang, C., Wang, Y., Li, X., Ren, L., Zhao, J., Hu, Y., Zhang, L., Fan, G., Xu, J., Gu, X., et al. (2020a). Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. Lancet.
10. Chan, J.F.W., Kok, K.H., Zhu, Z., Chu, H., To, K.K.W., Yuan, S., and Yuen, K.Y. (2020a). Genomic characterization of the 2019 novel human-pathogenic coronavirus isolated from a patient

with atypical pneumonia after visiting Wuhan. Emerg. Microbes Infect.

11. Wu, F., Zhao, S., Yu, B., Chen, Y.M., Wang, W., Song, Z.G., Hu, Y., Tao, Z.W., Tian, J.H., Pei, Y.Y., et al. (2020). A new coronavirus associated with human respiratory disease in China. Nature.
12. Zhou, P., Yang, X. Lou, Wang, X.G., Hu, B., Zhang, L., Zhang, W., Si, H.R., Zhu, Y., Li, B., Huang, C.L., et al. (2020). A pneumonia outbreak associated with a new coronavirus of probable bat origin. Nature.
13. Frieman, M., and Baric, R. (2008). Mechanisms of Severe Acute Respiratory Syndrome Pathogenesis and Innate Immunomodulation. Microbiol. Mol. Biol. Rev.
14. Harrison, A.G., Lin, T., and Wang, P. (2020). Mechanisms of SARS-CoV-2 Transmission and Pathogenesis. Trends Immunol.
15. Ackermann, M., Verleden, S.E., Kuehnel, M., Haverich, A., Welte, T., Laenger, F., Vanstapel, A., Werlein, C., Stark, H., Tzankov, A., et al. (2020). Pulmonary Vascular Endothelialitis, Thrombosis, and Angiogenesis in Covid-19. N. Engl. J. Med.
16. Borczuk, A.C., Salvatore, S.P., Seshan, S. V., Patel, S.S., Bussel, J.B., Mostyka, M., Elsoukkary, S., He, B., Del Vecchio, C., Fortarezza, F., et al. (2020). COVID-19 pulmonary pathology: a multi-institutional autopsy cohort from Italy and New York City. Mod. Pathol.
17. Galani, I.E., Rovina, N., Lampropoulou, V., Triantafyllia, V., Manioudaki, M., Pavlos, E., Koukaki, E., Fragkou, P.C., Panou, V., Rapti, V., et al. (2021). Untuned antiviral immunity in COVID-19 revealed by temporal type I/III interferon patterns and flu comparison. Nat. Immunol.
18. Hadjadj, J., Yatim, N., Barnabei, L., Corneau, A., Boussier, J., Smith, N., Péré, H., Charbit, B., Bondet, V., Chenevier-Gobeaux, C., et al. (2020). Impaired type I interferon activity and inflammatory responses in severe COVID-19 patients. Science (80-.).
19. Lucas, C., Wong, P., Klein, J., Castro, T.B.R., Silva, J., Sundaram, M., Ellingson, M.K., Mao, T., Oh, J.E., Israelow, B., et al. (2020). Longitudinal analyses reveal immunological misfiring in severe COVID-19. Nature.
20. Mathew, D., Giles, J.R., Baxter, A.E., Oldridge, D.A., Greenplate, A.R., Wu, J.E., Alanio, C., Kuri-Cervantes, L., Pampena, M.B., D'Andrea, K., et al. (2020). Deep immune profiling of COVID-19 patients reveals distinct immunotypes with therapeutic implications. Science (80-.).
21. Schulte-Schrepping, J., Reusch, N., Paclik, D., Baßler, K., Schlickeiser, S., Zhang, B., Krämer, B., Krammer, T., Brumhard, S., Bonaguro, L., et al. (2020). Severe COVID-19 Is Marked by a Dysregulated Myeloid Cell Compartment. Cell.
22. Stephenson, E., Reynolds, G., Botting, R.A., Calero-Nieto, F.J., Morgan, M., Tuong, Z.K., Bach, K., Sungnak, W., Worlock, K.B., Yoshida, M., et al. (2021). The cellular immune response to COVID-19 deciphered by single cell multi-omics across three UK centres. MedRxiv.

23. Su, Y., Chen, D., Yuan, D., Lausted, C., Choi, J., Dai, C.L., Voillet, V., Duvvuri, V.R., Scherler, K., Troisch, P., et al. (2020). Multi-Omics Resolves a Sharp Disease-State Shift between Mild and Moderate COVID-19. Cell.
24. Ren, X., Wen, W., Fan, X., Hou, W., Su, B., Cai, P., Li, J., Liu, Y., Tang, F., Zhang, F., et al. (2021). COVID-19 immune features revealed by a large-scale single cell transcriptome atlas. Cell.
25. Szabo, P.A., Dogra, P., Gray, J.I., Wells, S.B., Connors, T.J., Weisberg, S.P., Krupska, I., Matsumoto, R., Poon, M.M.L., Idzikowski, E., et al. (2020). Analysis of respiratory and systemic immune responses in COVID-19 reveals mechanisms of disease pathogenesis. MedRxiv.
26. Huang, N., Perez, P., Kato, T., Mikami, Y., Okuda, K., Gilmore, R.C., Conde, C.D., Gasmi, B., Stein, S., Beach, M., et al. (2020b). Integrated Single-Cell Atlases Reveal an Oral SARS-CoV-2 Infection and Transmission Axis. MedRxiv.
27. Lukassen, S., Chua, R.L., Trefzer, T., Kahn, N.C., Schneider, M.A., Muley, T., Winter, H., Meister, M., Veith, C., Boots, A.W., et al. (2020). SARS-CoV-2 receptor ACE2 and TMPRSS2 are predominantly expressed in a transient secretory cell type in subsegmental bronchial branches. BioRxiv.
28. Muus, C., Luecken, M.D., Eraslan, G., Waghray, A., Heimberg, G., Sikkema, L., Kobayashi, Y., Vaishnav, E.D., Subramanian, A., Smilie, C., et al. (2020). Integrated analyses of single-cell atlases reveal age, gender, and smoking status associations with cell type-specific expression of mediators of SARS-CoV-2 viral entry and highlights inflammatory programs in putative target cells. BioRxiv 2020.04.19.049254.
29. Sungnak, W., Huang, N., Bécavin, C., Berg, M., Queen, R., Litvinukova, M., Talavera-López, C., Maatz, H., Reichart, D., Sampaziotis, F., et al. (2020). SARS-CoV-2 entry factors are highly expressed in nasal epithelial cells together with innate immune genes. Nat. Med.
30. Ziegler, C.G.K., Allon, S.J., Nyquist, S.K., Mbano, I.M., Miao, V.N., Tzouanas, C.N., Cao, Y., Yousif, A.S., Bals, J., Hauser, B.M., et al. (2020). SARS-CoV-2 Receptor ACE2 Is an Interferon-Stimulated Gene in Human Airway Epithelial Cells and Is Detected in Specific Cell Subsets across Tissues. Cell.
31. Chua, R.L., Lukassen, S., Trump, S., Hennig, B.P., Wendisch, D., Pott, F., Debnath, O., Thürmann, L., Kurth, F., Völker, M.T., et al. (2020). COVID-19 severity correlates with airway epithelium–immune cell interactions identified by single-cell analysis. Nat. Biotechnol.
32. Hou, Y.J., Okuda, K., Edwards, C.E., Martinez, D.R., Asakura, T., Dinnon, K.H., Kato, T., Lee, R.E., Yount, B.L., Mascenik, T.M., et al. (2020). SARS-CoV-2 Reverse Genetics Reveals a Variable Infection Gradient in the Respiratory Tract. Cell.
33. Schaefer, I.M., Padera, R.F., Solomon, I.H., Kanjilal, S., Hammer, M.M., Hornick, J.L., and Sholl,

L.M. (2020). In situ detection of SARS-CoV-2 in lungs and airways of patients with COVID-19. Mod. Pathol.

34. Zhu, N., Wang, W., Liu, Z., Liang, C., Wang, W., Ye, F., Huang, B., Zhao, L., Wang, H., Zhou, W., et al. (2020). Morphogenesis and cytopathic effect of SARS-CoV-2 infection in human airway epithelial cells. Nat. Commun.
35. Blanco-Melo, D., Nilsson-Payant, B., Liu, W.-C., Moeller, R., Panis, M., Sachs, D., Albrecht, R., and tenOever, B.R. (2020). SARS-CoV-2 launches a unique transcriptional signature from in vitro, ex vivo, and in vivo systems. BioRxiv.
36. Ravindra, N.G., Alfajaro, M.M., Gasque, V., Habet, V., Wei, J., Filler, R.B., Huston, N.C., Wan, H., Szigeti-Buck, K., Wang, B., et al. (2020). Single-cell longitudinal analysis of SARS-CoV-2 infection in human airway epithelium. BioRxiv.
37. Chandrashekar, A., Liu, J., Martino, A.J., McMahan, K., Mercad, N.B., Peter, L., Tostanosk, L.H., Yu, J., Maliga, Z., Nekorchuk, M., et al. (2020). SARS-CoV-2 infection protects against rechallenge in rhesus macaques. Science (80-.).
38. Munster, V., Feldmann, F., Williamson, B., Doremalen, N. van, Lizzette Perez-Perez, Schultz, J., Meade-White, K., Okumura, A., Callison, J., Brumbaugh, B., et al. (2020). Respiratory disease and virus shedding in rhesus macaques inoculated with SARS-CoV-2. BioRxiv.
39. Speranza, E., Williamson, B.N., Feldmann, F., Sturdevant, G.L., Pérez, L.P., Meade-White, K., Smith, B.J., Lovaglio, J., Martens, C., Munster, V.J., et al. (2021). Single-cell RNA sequencing reveals SARS-CoV-2 infection dynamics in lungs of African green monkeys. Sci. Transl. Med.
40. Chan, J.F.W., Zhang, A.J., Yuan, S., Poon, V.K.M., Chan, C.C.S., Lee, A.C.Y., Chan, W.M., Fan, Z., Tsoi, H.W., Wen, L., et al. (2020b). Simulation of the Clinical and Pathological Manifestations of Coronavirus Disease 2019 (COVID-19) in a Golden Syrian Hamster Model: Implications for Disease Pathogenesis and Transmissibility. Clin. Infect. Dis.
41. Sia, S.F., Yan, L.M., Chin, A.W.H., Fung, K., Choy, K.T., Wong, A.Y.L., Kaewpreedee, P., Perera, R.A.P.M., Poon, L.L.M., Nicholls, J.M., et al. (2020). Pathogenesis and transmission of SARS-CoV-2 in golden hamsters. Nature.
42. Bao, L., Deng, W., Huang, B., Gao, H., Liu, J., Ren, L., Wei, Q., Yu, P., Xu, Y., Qi, F., et al. (2020). The pathogenicity of SARS-CoV-2 in hACE2 transgenic mice. Nature.
43. Israelow, B., Song, E., Mao, T., Lu, P., Meir, A., Liu, F., Alfajaro, M.M., Wei, J., Dong, H., Homer, R.J., et al. (2020). Mouse model of SARS-CoV-2 reveals inflammatory role of type i interferon signaling. J. Exp. Med.
44. Jiang, R. Di, Liu, M.Q., Chen, Y., Shan, C., Zhou, Y.W., Shen, X.R., Li, Q., Zhang, L., Zhu, Y., Si, H.R., et al. (2020). Pathogenesis of SARS-CoV-2 in Transgenic Mice Expressing Human

Angiotensin-Converting Enzyme 2. Cell.

45. Sun, S.H., Chen, Q., Gu, H.J., Yang, G., Wang, Y.X., Huang, X.Y., Liu, S.S., Zhang, N.N., Li, X.F., Xiong, R., et al. (2020). A Mouse Model of SARS-CoV-2 Infection and Pathogenesis. Cell Host Microbe.
46. Kim, Y. Il, Kim, S.G., Kim, S.M., Kim, E.H., Park, S.J., Yu, K.M., Chang, J.H., Kim, E.J., Lee, S., Casel, M.A.B., et al. (2020b). Infection and Rapid Transmission of SARS-CoV-2 in Ferrets. Cell Host Microbe.
47. Richard, M., Kok, A., de Meulder, D., Bestebroer, T.M., Lamers, M.M., Okba, N.M.A., Fentener van Vlissingen, M., Rockx, B., Haagmans, B.L., Koopmans, M.P.G., et al. (2020). SARS-CoV-2 is transmitted via contact and via the air between ferrets. Nat. Commun.
48. Muñoz-Fontela, C., Dowling, W.E., Funnell, S.G.P., Gsell, P.S., Riveros-Balta, A.X., Albrecht, R.A., Andersen, H., Baric, R.S., Carroll, M.W., Cavaleri, M., et al. (2020). Animal models for COVID-19. Nature.
49. Bastard, P., Rosen, L.B., Zhang, Q., Michailidis, E., Hoffmann, H.H., Zhang, Y., Dorgham, K., Philippot, Q., Rosain, J., Béziat, V., et al. (2020). Autoantibodies against type I IFNs in patients with life-threatening COVID-19. Science (80-.).
50. Combes, A.J., Courau, T., Kuhn, N.F., Hu, K.H., Ray, A., Chen, W.S., Chew, N.W., Cleary, S.J., Kushnoor, D., Reeder, G.C., et al. (2021). Global absence and targeting of protective immune states in severe COVID-19. Nature.
51. Zhang, Q., Liu, Z., Moncada-Velez, M., Chen, J., Ogishi, M., Bigio, B., Yang, R., Arias, A.A., Zhou, Q., Han, J.E., et al. (2020). Inborn errors of type I IFN immunity in patients with life-threatening COVID-19. Science (80-.).
52. Gierahn, T.M., Wadsworth, M.H., Hughes, T.K., Bryson, B.D., Butler, A., Satija, R., Fortune, S., Christopher Love, J., and Shalek, A.K. (2017). Seq-Well: Portable, low-cost rna sequencing of single cells at high throughput. Nat. Methods.
53. Hughes, T.K., Wadsworth, M.H., Gierahn, T.M., Do, T., Weiss, D., Andrade, P.R., Ma, F., Silva, B.J. de A., Shao, S., Tsoi, L.C., et al. (2019). Highly Efficient, Massively-Parallel Single-Cell RNA-Seq Reveals Cellular States and Molecular Features of Human Skin Pathology. BioRxiv.
54. Deprez, M., Zaragosi, L.E., Truchi, M., Becavin, C., García, S.R., Arguel, M.J., Plaisant, M., Magnone, V., Lebrigand, K., Abelanet, S., et al. (2020). A single-cell atlas of the human healthy airways. Am. J. Respir. Crit. Care Med.
55. Garcıá, S.R., Deprez, M., Lebrigand, K., Cavard, A., Paquet, A., Arguel, M.J., Magnone, V., Truchi, M., Caballero, I., Leroy, S., et al. (2019). Novel dynamics of human mucociliary differentiation revealed by single-cell RNA sequencing of nasal epithelial cultures. Dev.

56. Ordovas-Montanes, J., Dwyer, D.F., Nyquist, S.K., Buchheit, K.M., Vukovic, M., Deb, C., Wadsworth, M.H., Hughes, T.K., Kazer, S.W., Yoshimoto, E., et al. (2018). Allergic inflammatory memory in human respiratory epithelial progenitor cells. Nature.
57. Montoro, D.T., Haber, A.L., Biton, M., Vinarsky, V., Lin, B., Birket, S.E., Yuan, F., Chen, S., Leung, H.M., Villoria, J., et al. (2018). A revised airway epithelial hierarchy includes CFTR-expressing ionocytes. Nature.
58. Plasschaert, L.W., Žilionis, R., Choo-Wing, R., Savova, V., Knehr, J., Roma, G., Klein, A.M., and Jaffe, A.B. (2018). A single-cell atlas of the airway epithelium reveals the CFTR-rich pulmonary ionocyte. Nature.
59. Basak, O., Beumer, J., Wiebrands, K., Seno, H., van Oudenaarden, A., and Clevers, H. (2017). Induced Quiescence of Lgr5+ Stem Cells in Intestinal Organoids Enables Differentiation of Hormone-Producing Enteroendocrine Cells. Cell Stem Cell.
60. Hoffmann, M., Kleine-Weber, H., Schroeder, S., Krüger, N., Herrler, T., Erichsen, S., Schiergens, T.S., Herrler, G., Wu, N.H., Nitsche, A., et al. (2020). SARS-CoV-2 Cell Entry Depends on ACE2 and TMPRSS2 and Is Blocked by a Clinically Proven Protease Inhibitor. Cell.
61. Li, W., Moore, M.J., Vasllieva, N., Sui, J., Wong, S.K., Berne, M.A., Somasundaran, M., Sullivan, J.L., Luzuriaga, K., Greeneugh, T.C., et al. (2003). Angiotensin-converting enzyme 2 is a functional receptor for the SARS coronavirus. Nature.
62. Wang, Q., Zhang, Y., Wu, L., Niu, S., Song, C., Zhang, Z., Lu, G., Qiao, C., Hu, Y., Yuen, K.Y., et al. (2020a). Structural and Functional Basis of SARS-CoV-2 Entry by Using Human ACE2. Cell.
63. Wrapp, D., Wang, N., Corbett, K.S., Goldsmith, J.A., Hsieh, C.L., Abiona, O., Graham, B.S., and McLellan, J.S. (2020). Cryo-EM structure of the 2019-nCoV spike in the prefusion conformation. Science (80-.).
64. Yan, R., Zhang, Y., Li, Y., Xia, L., Guo, Y., and Zhou, Q. (2020). Structural basis for the recognition of SARS-CoV-2 by full-length human ACE2. Science (80-.).
65. Raj, V.S., Mou, H., Smits, S.L., Dekkers, D.H.W., Müller, M.A., Dijkman, R., Muth, D., Demmers, J.A.A., Zaki, A., Fouchier, R.A.M., et al. (2013). Dipeptidyl peptidase 4 is a functional receptor for the emerging human coronavirus-EMC. Nature.
66. Yeager, C.L., Ashmun, R.A., Williams, R.K., Cardellichio, C.B., Shapiro, L.H., Look, A.T., and Holmes, K. V. (1992). Human aminopeptidase N is a receptor for human coronavirus 229E. Nature.
67. Bochkov, Y.A., Watters, K., Ashraf, S., Griggs, T.F., Devries, M.K., Jackson, D.J., Palmenberg, A.C., and Gern, J.E. (2015). Cadherin-related family member 3, a childhood asthma susceptibility gene product, mediates rhinovirus C binding and replication. Proc. Natl. Acad. Sci. U. S. A.
68. Bergen, V., Lange, M., Peidli, S., Wolf, F.A., and Theis, F.J. (2020). Generalizing RNA velocity

to transient cell states through dynamical modeling. Nat. Biotechnol.

69. La Manno, G., Soldatov, R., Zeisel, A., Braun, E., Hochgerner, H., Petukhov, V., Lidschreiber, K., Kastriti, M.E., Lönnerberg, P., Furlan, A., et al. (2018). RNA velocity of single cells. Nature.
70. Tata, P.R., Mou, H., Pardo-Saganta, A., Zhao, R., Prabhu, M., Law, B.M., Vinarsky, V., Cho, J.L., Breton, S., Sahay, A., et al. (2013). Dedifferentiation of committed epithelial cells into stem cells in vivo. Nature.
71. Blume, C., Jackson, C.L., Spalluto, C.M., Legebeke, J., Nazlamova, L., Conforti, F., Perotin, J.M., Frank, M., Butler, J., Crispin, M., et al. (2021). A novel ACE2 isoform is expressed in human respiratory epithelia and is upregulated in response to interferons and RNA respiratory virus infection. Nat. Genet.
72. Ng, K.W., Attig, J., Bolland, W., Young, G.R., Major, J., Wrobel, A.G., Gamblin, S., Wack, A., and Kassiotis, G. (2020). Tissue-specific and interferon-inducible expression of nonfunctional ACE2 through endogenous retroelement co-option. Nat. Genet.
73. Onabajo, O.O., Banday, A.R., Stanifer, M.L., Yan, W., Obajemu, A., Santer, D.M., Florez-Vargas, O., Piontkivska, H., Vargas, J.M., Ring, T.J., et al. (2020). Interferons and viruses induce a novel truncated ACE2 isoform and not the full-length SARS-CoV-2 receptor. Nat. Genet.
74. Cao, Y., Guo, Z., Vangala, P., Donnard, E., Liu, P., McDonel, P., Ordovas-Montanes, J., Shalek, A.K., Finberg, R.W., Wang, J.P., et al. (2020). Single-cell analysis of upper airway cells reveals host-viral dynamics in influenza infected adults. BioRxiv.
75. Lemieux, J.E., Siddle, K.J., Shaw, B.M., Loreth, C., Schaffner, S.F., Gladden-Young, A., Adams, G., Fink, T., Tomkins-Tinch, C.H., Krasilnikova, L.A., et al. (2020). Phylogenetic analysis of SARS-CoV-2 in Boston highlights the impact of superspreading events. Science (80-.).
76. Wood, D.E., Lu, J., and Langmead, B. (2019). Improved metagenomic analysis with Kraken 2. Genome Biol.
77. Kim, D., Lee, J.Y., Yang, J.S., Kim, J.W., Kim, V.N., and Chang, H. (2020a). The Architecture of SARS-CoV-2 Transcriptome. Cell.
78. Fleming, S.J., Marioni, J.C., and Babadi, M. (2019). CellBender remove-background: A deep generative model for unsupervised removal of background noise from scRNA-seq datasets. BioRxiv.
79. Kotliar, D., Lin, A.E., Logue, J., Hughes, T.K., Khoury, N.M., Raju, S.S., Wadsworth, M.H., Chen, H., Kurtz, J.R., Dighero-Kemp, B., et al. (2020). Single-Cell Profiling of Ebola Virus Disease In Vivo Reveals Viral and Host Dynamics. Cell.
80. Fung, T.S., and Liu, D.X. (2019). Human Coronavirus: Host-Pathogen Interaction. Annu. Rev. Microbiol.

81. Hu, B., Guo, H., Zhou, P., and Shi, Z.L. (2020). Characteristics of SARS-CoV-2 and COVID-19. Nat. Rev. Microbiol.
82. Sawicki, S.G., Sawicki, D.L., and Siddell, S.G. (2007). A Contemporary View of Coronavirus Transcription. J. Virol.
83. Krähling, V., Stein, D.A., Spiegel, M., Weber, F., and Mühlberger, E. (2009). Severe Acute Respiratory Syndrome Coronavirus Triggers Apoptosis via Protein Kinase R but Is Resistant to Its Antiviral Activity. J. Virol.
84. Zhao, X., Guo, F., Liu, F., Cuconati, A., Chang, J., Block, T.M., and Guo, J.T. (2014). Interferon induction of IFITM proteins promotes infection by human coronavirus OC43. Proc. Natl. Acad. Sci. U. S. A.
85. Daniloski, Z., Jordan, T.X., Wessels, H.H., Hoagland, D.A., Kasela, S., Legut, M., Maniatis, S., Mimitou, E.P., Lu, L., Geller, E., et al. (2021). Identification of Required Host Factors for SARS-CoV-2 Infection in Human Cells. Cell.
86. Wang, R., Simoneau, C.R., Kulsuptrakul, J., Bouhaddou, M., Travisano, K.A., Hayashi, J.M., Carlson-Stevermer, J., Zengel, J.R., Richards, C.M., Fozouni, P., et al. (2021). Genetic Screens Identify Host Factors for SARS-CoV-2 and Common Cold Coronaviruses. Cell.
87. Schneider, W.M., Luna, J.M., Hoffmann, H.H., Sánchez-Rivera, F.J., Leal, A.A., Ashbrook, A.W., Le Pen, J., Ricardo-Lax, I., Michailidis, E., Peace, A., et al. (2021). Genome-Scale Identification of SARS-CoV-2 and Pan-coronavirus Host Factor Networks. Cell.
88. Wei, J., Alfajaro, M.M., DeWeirdt, P.C., Hanna, R.E., Lu-Culligan, W.J., Cai, W.L., Strine, M.S., Zhang, S.M., Graziano, V.R., Schmitz, C.O., et al. (2021). Genome-wide CRISPR Screens Reveal Host Factors Critical for SARS-CoV-2 Infection. Cell.
89. Kamitani, W., Huang, C., Narayanan, K., Lokugamage, K.G., and Makino, S. (2009). A two-pronged strategy to suppress host protein synthesis by SARS coronavirus Nsp1 protein. Nat. Struct. Mol. Biol.
90. Lokugamage, K.G., Narayanan, K., Nakagawa, K., Terasaki, K., Ramirez, S.I., Tseng, C.-T.K., and Makino, S. (2015). Middle East Respiratory Syndrome Coronavirus nsp1 Inhibits Host Gene Expression by Selectively Targeting mRNAs Transcribed in the Nucleus while Sparing mRNAs of Cytoplasmic Origin. J. Virol.
91. Knoops, K., Kikkert, M., Van Den Worm, S.H.E., Zevenhoven-Dobbe, J.C., Van Der Meer, Y., Koster, A.J., Mommaas, A.M., and Snijder, E.J. (2008). SARS-coronavirus replication is supported by a reticulovesicular network of modified endoplasmic reticulum. PLoS Biol.
92. Menachery, V.D., Eisfeld, A.J., Schäfer, A., Josset, L., Sims, A.C., Proll, S., Fan, S., Li, C., Neumann, G., Tilton, S.C., et al. (2014). Pathogenic influenza viruses and coronaviruses utilize

similar and contrasting approaches to control interferon-stimulated gene responses. MBio.

93. Banerjee, A.K., Blanco, M.R., Bruce, E.A., Honson, D.D., Chen, L.M., Chow, A., Bhat, P., Ollikainen, N., Quinodoz, S.A., Loney, C., et al. (2020). SARS-CoV-2 Disrupts Splicing, Translation, and Protein Trafficking to Suppress Host Defenses. Cell.
94. Konno, Y., Kimura, I., Uriu, K., Fukushi, M., Irie, T., Koyanagi, Y., Sauter, D., Gifford, R.J., Nakagawa, S., and Sato, K. (2020). SARS-CoV-2 ORF3b Is a Potent Interferon Antagonist Whose Activity Is Increased by a Naturally Occurring Elongation Variant. Cell Rep.
95. Snijder, E.J., Limpens, R.W.A.L., de Wilde, A.H., de Jong, A.W.M., Zevenhoven-Dobbe, J.C., Maier, H.J., Faas, F.F.G.A., Koster, A.J., and Bárcena, M. (2020). A unifying structural and functional model of the coronavirus replication organelle: Tracking down RNA synthesis. PLoS Biol.
96. Broggi, A., Granucci, F., and Zanoni, I. (2020). Type III interferons: Balancing tissue tolerance and resistance to pathogen invasion. J. Exp. Med.
97. Silvin, A., Chapuis, N., Dunsmore, G., Goubet, A.G., Dubuisson, A., Derosa, L., Almire, C., Hénon, C., Kosmider, O., Droin, N., et al. (2020). Elevated Calprotectin and Abnormal Myeloid Cell Subsets Discriminate Severe from Mild COVID-19. Cell.
98. Bruchez, A., Sha, K., Johnson, J., Chen, L., Stefani, C., McConnell, H., Gaucherand, L., Prins, R., Matreyek, K.A., Hume, A.J., et al. (2020). MHC class II transactivator CIITA induces cell resistance to ebola virus and SARS-like coronaviruses. Science (80-.).
99. Tang, Y., Ziegler, C.G.K., Miao, V.N., Navia, A.W., Bromley, J.D., Wilson, K.J., Pride, Y., Hasan, M., Christian, T., Laird, H., et al. (2020). Human Nasopharyngeal Swab Processing for Viable Single-Cell Suspension. Protocols.Io.
100. Aicher, T.P., Carroll, S., Raddi, G., Gierahn, T., Wadsworth, M.H., Hughes, T.K., Love, C., and Shalek, A.K. (2019). Seq-Well: A sample-efficient, portable picowell platform for massively parallel single-cell RNA sequencing. In Methods in Molecular Biology, p.
101. Li, B., Gould, J., Yang, Y., Sarkizova, S., Tabaka, M., Ashenberg, O., Rosen, Y., Slyper, M., Kowalczyk, M.S., Villani, A.C., et al. (2020). Cumulus provides cloud-based data analysis for large-scale single-cell and single-nucleus RNA-seq. Nat. Methods.
102. Macosko, E.Z., Basu, A., Satija, R., Nemesh, J., Shekhar, K., Goldman, M., Tirosh, I., Bialas, A.R., Kamitaki, N., Martersteck, E.M., et al. (2015). Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. Cell.
103. Petukhov, V., Guo, J., Baryawno, N., Severe, N., Scadden, D.T., Samsonova, M.G., and Kharchenko, P. V. (2018). dropEst: Pipeline for accurate estimation of molecular counts in droplet-based single-cell RNA-seq experiments. Genome Biol.

104. Stuart, T., Butler, A., Hoffman, P., Hafemeister, C., Papalexi, E., Mauck, W.M., Hao, Y., Stoeckius, M., Smibert, P., and Satija, R. (2019). Comprehensive Integration of Single-Cell Data. Cell.
105. McDavid, A., Finak, G., Chattopadyay, P.K., Dominguez, M., Lamoreaux, L., Ma, S.S., Roederer, M., and Gottardo, R. (2013). Data exploration, quality control and testing in single-cell qPCR-based gene expression experiments. Bioinformatics.
106. Love, M.I., Anders, S., and Huber, W. (2014). Differential analysis of count data - the DESeq2 package.
107. Huang, D.W., Sherman, B.T., and Lempicki, R.A. (2009). Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat. Protoc.
108. Korotkevich, G., Sukhov, V., Budin, N., Shpak, B., Artyomov, M.N., and Sergushichev, A. (2021). Fast gene set enrichment analysis. BioRxiv.
109. R Core Team (2019). R: A language and environment for statistical computing. R Found. Stat. Comput.
110. Wickham, H. (2016). ggplot2: Elegant Graphics for Data Analysis.
111. Butler, A., Hoffman, P., Smibert, P., Papalexi, E., and Satija, R. (2018). Integrating single-cell transcriptomic data across different conditions, technologies, and species. Nat. Biotechnol.
112. Gu, Z., Eils, R., and Schlesner, M. (2016). Complex heatmaps reveal patterns and correlations in multidimensional genomic data. Bioinformatics.
113. Gu, Z., Gu, L., Eils, R., Schlesner, M., and Brors, B. (2014). Circlize implements and enhances circular visualization in R. Bioinformatics.
114. Ballestar, E., Farber, D.L., Glover, S., Horwitz, B., Meyer, K., Nikolić, M., Ordovas-Montanes, J., Sims, P., Shalek, A., Vandamme, N., et al. (2020). Single cell profiling of COVID-19 patients: An international data resource from multiple tissues. MedRxiv.

Chapter 3: The tumor microenvironment drives transcriptional phenotypes and their plasticity in metastatic pancreatic cancer

Srivatsan Raghavan*, Peter S. Winter *, Andrew W. Navia*, Hannah L. Williams*, Alan DenAdel, Radha L. Kalekar, Jennyfer Galvez-Reyes, Kristen E. Lowder, Nolawit Mulugeta, Manisha S. Raghavan, Ashir A. Borah, Kevin S. Kapner, Sara A. Väyrynen, Andressa Dias Costa, Raymond W.S. Ng, Junning Wang, Emma Reilly, Dorisanne Y. Ragon, Lauren K. Brais, Alex M. Jaeger, Liam F. Spurr, Yvonne Y. Li, Andrew D. Cherniack , Isaac Wakiro, Asaf Rotem , Bruce E. Johnson , James M. McFarland, Ewa T. Sicinska, Tyler E. Jacks, Thomas E. Clancy, Kimberly Perez, Douglas A. Rubinson, Kimmie Ng, James M. Cleary, Lorin Crawford, Scott R. Manalis, Jonathan A. Nowak, Brian M. Wolpin#, William C. Hahn#, Andrew J. Aguirre#, Alex K. Shalek#

*These authors contributed equally to this work
#These authors contributed equally to this work

3.1 Abstract

Bulk transcriptomic studies have defined classical and basal-like gene expression subtypes in pancreatic ductal adenocarcinoma (PDAC) that correlate with survival and response to chemotherapy; however, the underlying mechanisms that govern these subtypes and their heterogeneity remain elusive. Here, we performed single-cell RNA-sequencing of 23 metastatic PDAC needle biopsies and matched organoid models to understand how tumor cell-intrinsic features and extrinsic factors in the tumor microenvironment (TME) shape PDAC cancer cell phenotypes. We identify a novel cancer cell state that co-expresses basal-like and classical signatures, demonstrates upregulation of developmental and KRAS-driven gene expression programs, and represents a transitional intermediate between the basal-like and classical poles. Further, we observe structure to the metastatic TME supporting a model whereby reciprocal intercellular signaling shapes the local microenvironment and influences cancer cell transcriptional subtypes. In organoid culture, we find that transcriptional phenotypes are plastic and strongly skew toward the classical expression state, irrespective of genotype. Moreover, we show that patient-relevant transcriptional heterogeneity can be rescued by supplementing organoid media with factors found in the TME in a subtype-specific manner. Collectively, our study demonstrates that

distinct microenvironmental signals are critical regulators of clinically relevant PDAC transcriptional states and their plasticity, identifies the necessity for considering the TME in cancer modeling efforts, and provides a generalizable approach for delineating the cell-intrinsic versus -extrinsic factors that govern tumor cell phenotypes.

3.2 Introduction

Classification of human malignancies by genotype has provided important insights into tumor biology as well as a framework to guide therapeutic selection in many cancers[1]. However, tumors also exhibit clinically relevant transcriptional variation that can influence malignant progression and therapeutic response. The application of single-cell RNA-sequencing (scRNA-seq) to tumor specimens has afforded a means to characterize the malignant and non-malignant cellular components of the tumor microenvironment (TME) and their heterogeneity at unprecedented resolution[2-9]. These analytical approaches have further enabled the re-examination of existing transcriptional taxonomies, revealing structured heterogeneity within malignant populations and reframing our understanding of bulk measurements in multiple cancers[3,9,10-13].

The phenotypic variability observed in human tumors often reflects the underlying cancer cell genetics. Specific mutations can program cancer cell states and, in some cases, serve as biomarkers for treatment[8-11]. Yet, in other instances, transcriptional phenotypes are not strongly associated with specific mutational patterns[14]. In these tumors, cell-extrinsic TME interactions may influence malignant cellular attributes, but our understanding of reciprocal signaling between malignant cells and the TME is rudimentary. Mapping the cell-intrinsic and -extrinsic factors that impact tumor cell states and determining which ones drive phenotypic transitions would yield important insights into the biologic basis for clinical disease phenotypes and drug resistance.

For pancreatic ductal adenocarcinoma (PDAC), bulk RNA-seq profiling has defined two major transcriptional programs, basal-like/squamous (hereafter referred to as basal) and classical. The basal subtype is strongly associated with a poorer prognosis and greater treatment resistance[15-25], but the roles of cell-intrinsic and -extrinsic factors in determining these cell states and their sensitivity to different therapies are not well understood. A limited number of genomic alterations, including *TP53* mutational status and *c-MYC* or *KRAS* amplifications, have been associated with the more therapy-resistant basal state[17-19,26,27]. Recent studies have also suggested that high levels of KRAS expression and signaling can induce the basal state, but others have demonstrated that

basal PDAC cells exhibit RAS-independence[19,20,28,29]. These findings suggest that while genomic activation of *KRAS* plays an important role in oncogenesis, other non-genetic and microenvironmental factors may also be critical in regulating downstream cellular states.

Although the majority of patients with PDAC present with and succumb to metastatic disease[30], our current understanding of PDAC is largely derived from resected primary tumors[15,18,30]. While several recent studies have described the desmoplastic stromal microenvironment and immune infiltration in primary PDAC[31-36], we lack a detailed characterization of the immune and stromal cells that constitute metastatic PDAC lesions. The local TME in the pancreas is likely different from metastatic sites in other organs[37], and given the strong association of transcriptional subtype with survival and drug resistance[15-25], understanding whether specific inputs from the metastatic niche can specify transcriptional phenotype is of great importance to targeting therapeutic resistance in PDAC.

To better understand the interplay between genetics, transcriptional state, and the metastatic TME, we developed and employed an optimized translational workflow to perform both high-resolution profiling of PDAC patient tissue using scRNA-seq[38,39] and derivation of matched organoid models[26,40] from the same metastatic core needle biopsy. Using matched *in vivo* observations and *ex vivo* experimental studies, we describe a tumor cell atlas of metastatic PDAC, identify a new intermediate transitional PDAC cancer cell state, uncover distinct site- and subtype-specific TMEs, and demonstrate that microenvironmental signals are critical regulators of transcriptional subtypes and their plasticity.

3.3 Results

3.3.1 A clinical pipeline for matched single-cell profiling and organoid model generation

We established a pipeline for collecting core needle biopsies from patients with metastatic PDAC (n=23) to generate matched scRNA-seq profiles and organoid models (**Figure 3.1A**; **Supplemental Figure S3.1A**; **Supplemental Table S3.1**). Most samples were obtained from metastatic lesions residing in the liver (19/23), and the majority (21/23) were analyzed by targeted DNA-sequencing, yielding the expected mutational pattern for this disease (**Figure 3.1B**) [15,17,18].

Our pipeline generated approximately 1,000 high-quality single cells per biopsy (n=23,042 total cells) and successful early-passage organoid cultures from 70% of patient tumor samples (16/23

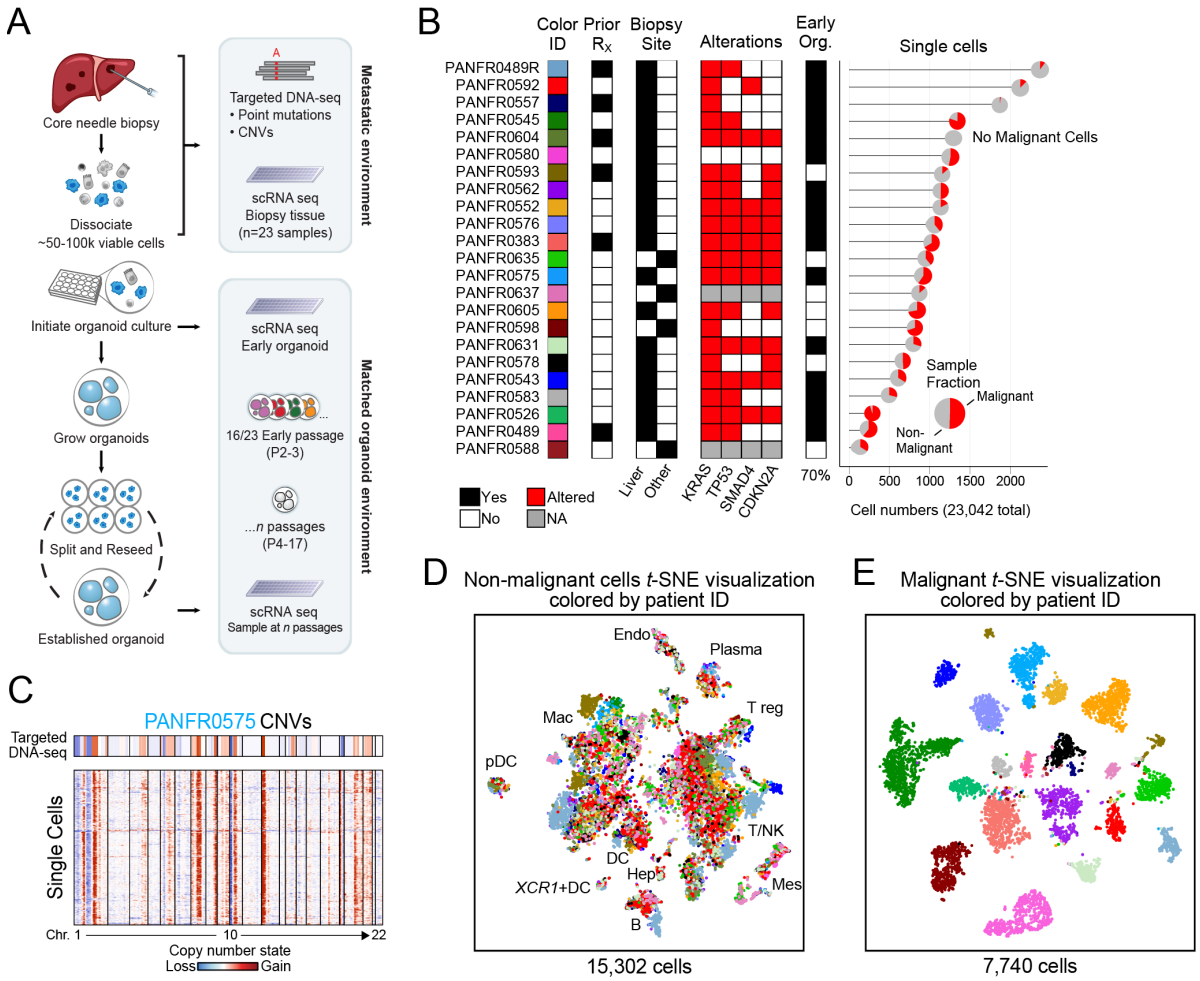


Figure 3.1: A clinical pipeline for matched single-cell RNA-seq and organoid generation from metastatic PDAC biopsies.

(A) Pipeline for collecting patient samples, and dissociation and allocation for scRNA-seq and parallel organoid development. **(B)** Clinical and molecular features for all patients included in the dataset (R_x = Therapy; Other = Adrenal (PANFR0637), Omentum (PANFR0635, PANFR0598), Peritoneum (PANFR0588); Org. at P2 = Organoid measured at passage 2). Mutations were determined by bulk targeted DNA-seq (Red, Altered; White, wildtype; Grey, Data not available). Number of single cells captured per biopsy and their malignant and non-malignant fraction is visualized at the right. **(C)** Example bulk targeted DNA-seq (top) and single-cell inferred CNV profiles (rows, bottom) arranged by chromosome (columns) from PANFR0575. **(D-E)** *t*-distributed stochastic neighbor embedding (*t*-SNE) visualization for non-malignant (**D**) and malignant (**E**) single cells in the biopsy cohort. Cells are colored by patient as in **B**. Endo, Endothelial; Mes, Mesenchymal; B, B-cell; Hep, Hepatocyte; DC, Dendritic cell; pDC, Plasmacytoid dendritic cell; Mac, Macrophage; T, T-cell; NK, Natural killer cell.

samples reaching at least passage 2; **Figure 3.1B**; **Supplemental Figure 3.1A**,**B**). Dimensionality reduction and shared nearest neighbor (SNN) clustering of the biopsy cells revealed substantial heterogeneity at the single-cell level (**Supplemental Figure S3.1C**,**D**; **Methods**). Consistent with other studies of human cancer, we observed patient-specific and admixed clusters of single cells suggesting the presence of both malignant and non-malignant cells in each biopsy[2,4,5,7,8]. To confirm which clusters were comprised of malignant cells, we inferred transcriptome-wide CNVs from our single-cell data as previously described[3,13]. CNV alteration scores separated putative cancerous and non-cancerous cells in each biopsy and demonstrated high concordance with reference targeted DNA-seq (**Figure 3.1C**; **Supplemental Figure S3.1E**,**F**). CNV analysis paired with manual inspection of expression patterns for known markers across single cells supported the identification of cancerous cells as well as 11 unique non-cancerous cell types (**Figure 3.1D**,**E**; **Supplemental Figure S3.1D-I**; **Supplemental Table S3.2**). Thus, we established a robust workflow capable of recovering high quality malignant (n=7,740) and non-malignant (n=15,302) populations from metastatic PDAC needle biopsies while also enabling simultaneous generation of matched organoid models.

3.3.2 Tumor cell transcriptional subtypes in metastatic PDAC include an intermediate transitional state

We applied principal component analysis (PCA) to examine transcriptional variation across cancer cells from all biopsy samples. CNV-altered cells from one biopsy, PANFR0580, separated from the rest of the samples (**Figure 3.1B**; **Supplemental Figure S3.2A**). Based on expression of known neuroendocrine markers (*TTR, CHGA*) and subsequent pathology review we reclassified this sample as a pancreatic neuroendocrine tumor (PanNET) and used it as a non-PDAC reference cell population. Among the remaining 7,078 PDAC cells, we found that genes driving the first 3 PCs were enriched for signatures of epithelial/mesenchymal transition [EMT, PC1[41]], basal/classical state [PC2[22]], and cell cycle [PC3[7]] (**Supplemental Figure S3.2B**). When we scored all malignant cells within our cohort for basal and classical gene expression, we observed that they inhabited a graded continuum of expression states from strongly basal to strongly classical (**Figure 3.2A**). Correlation analysis across malignant cells revealed 1,909 genes significantly associated with either basal or classical expression scores (**Supplemental Figure S3.2C**; **Supplemental Table S3.3**; **Methods**). Inspection of these genes revealed that basal cells

are defined by squamous and mesenchymal features and co-express programs associated with transforming growth factor beta (TGF-β) signaling, WNT signaling, and cell cycle progression[2,7,41]. Conversely, epithelial and pancreatic lineage programs are enriched in classical subtype PDAC cells (**Supplemental Figure S3.2D**,**E**).

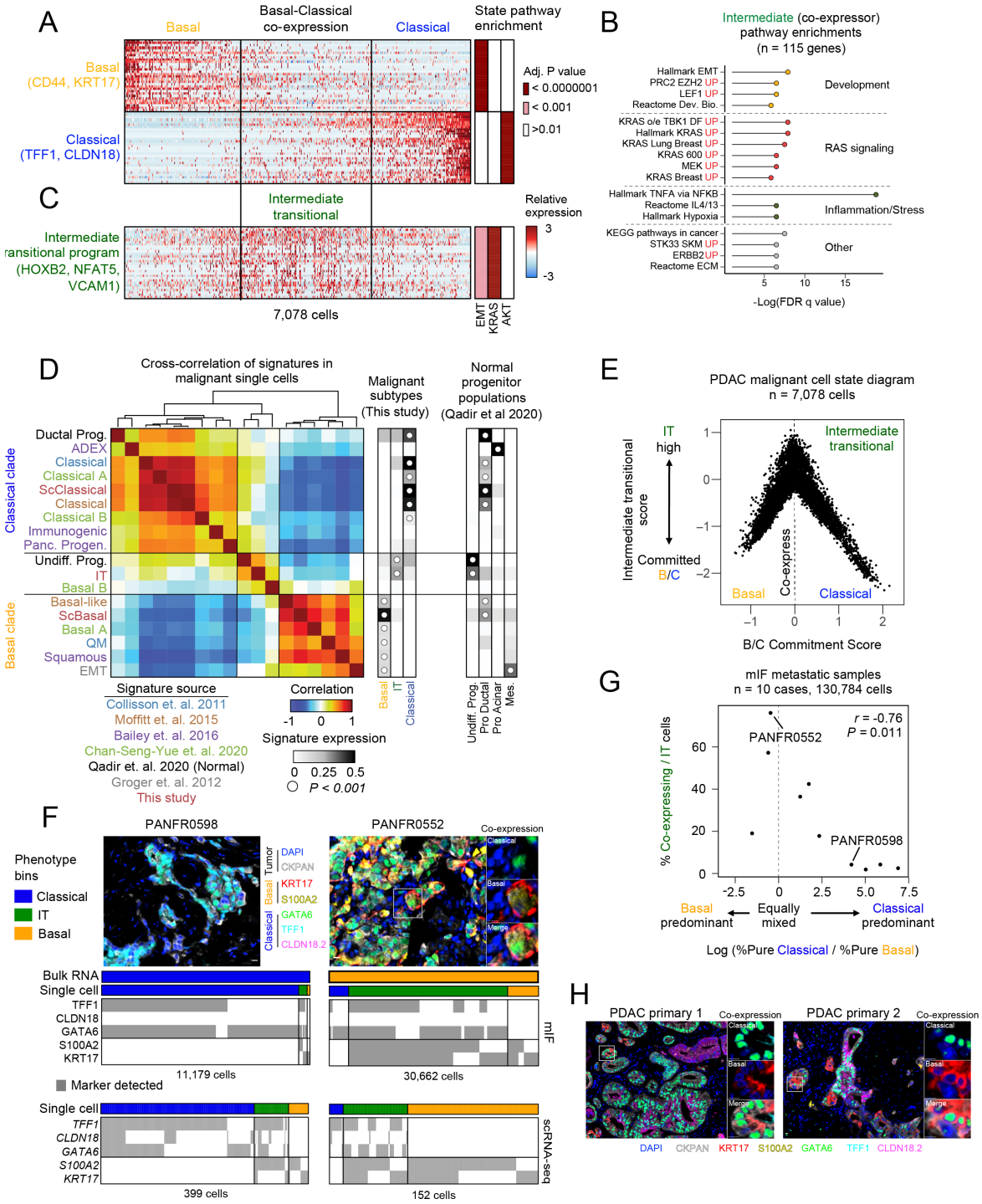


Figure 3.2: An intermediate transitional state bridges basal and classical phenotypes.

(A) Heatmaps depict the expression of basal and classical expression programs and highlight the co-expressing intermediates (n=30 genes each). **(B)** Gene set enrichment analysis for the 115 genes specific to the co-expressing intermediate state. **(C)** The intermediate transitional ("IT") expression program (n = 30 genes) is enriched by co-expressing cells. Enrichment adjusted *P*-values (hypergeometric test) for EMT, *KRAS*, and *AKT* gene sets are indicated at right for each gene expression program in **A** and **C**. **(D)** Cross-correlation between new and previously proposed expression signatures (rows and columns; text color = source, below) in our PDAC single-cells. Average expression for each signature (rows) is shown at the right for cells in the malignant subtypes from our cohort and the normal pancreatic progenitor cells from Qadir et al., 2020. White dot indicates the subset with the highest average significant expression for each signature (Kruskal-Wallis test); no white dot indicates no significant expression. **(E)** Malignant cell state diagram for PDAC. Basal-classical commitment score (x axis) and IT score (y axis) for all 7,078 malignant cells (**Methods**). **(F)** Multiplex immunofluorescence analysis (mIF) identifies co-expressing IT cells in matched metastatic samples. Top are representative images from two cases (white box indicates region for co-expression insets at right), and bottom indicates marker detection patterns for mIF and matched scRNA-seq data (**Methods**). Scale bar represents 10 μm. **(G)** Frequency of co-expressing IT cells is correlated with balanced representation of pure basal and classical phenotypes by mIF within individual samples. Log ratio of % basal and classical cells in each sample (x axis) versus their % co-expressing / IT cells (y axis). **(H)** Co-expressing IT cells are also identified in primary PDAC samples by mIF. Scale bar represents 10 μm.

Strikingly, we observed that the basal and classical programs were not mutually exclusive; rather, we identified a large population of cells that co-expressed features of both programs to varying degrees (**Figure 3.2A**; **Supplemental Figure S3.3A**,**B**). In developmental contexts, cell state commitment is often a continuous process where mixing/co-expression of state markers indicates state transitions[14]. Similarly, the large fraction of intermediate co-expressing cells identified in our single-cell snapshots suggests state transitions may be an ongoing and frequent process in human PDAC tumors. We identified 115 genes whose expression was correlated with this co-expressor intermediate state and enriched for developmental, Ras signaling, and inflammation/stress response gene sets (**Figure 3.2B**; **Supplemental Figure S3.3C**,**D**; **Supplemental Table S3.3**; **Methods**). Signatures of RAS signaling were enriched in the intermediate state even compared with basal and classical programs, and, by contrast, classical phenotypes were enriched for Akt-associated gene sets and showed little evidence of EMT or RAS enrichments (**Figure 3.2A-B**; **Supplemental Figure S3.2E**).

Since this intermediate signature showed enrichment for developmental gene programs, we next assessed whether this signature overlapped with any phenotypes recently reported in the normal pancreas progenitor niche[42]. We found that both basal and classical gene expression

signatures were expressed by pro-ductal progenitor cells, while the intermediate gene expression program was enriched in an undifferentiated, stress-responsive progenitor population (**Supplemental Figure S3.3E**,**F**)[42]. Thus, based on its enrichment for developmental and stress-responsive gene sets, overlap with populations in the normal progenitor niche, and co-expression of basal and classical programs suggestive of a transitional state, we termed this phenotype "Intermediate transitional" (IT) (**Figure 3.2C**).

To further contextualize this cell state, we compared signatures proposed by prior bulk RNA-sequencing studies to clarify potential inter-relationships[15,17,19,20,22]. Pairwise correlation of all established signatures in malignant cells revealed that many contribute overlapping information and reflect similar underlying biology within either basal or classical clades, but that the IT signature is unique and not well described by established bulk RNA-seq signatures (**Figure 3.2D**). Taken together, these findings suggest that malignant PDAC cells organize in a tripartite cell state framework that spans committed basal and classical phenotypes, with considerable signature co-expression in single cells (**Figure 3.2E**). Similar to the variation in EMT scores observed in basal tumor cells (**Supplemental Figure S3.3A**)[19,21], we noted heterogeneity among co-expressing cells for the IT program.

3.3.3 Multiplex immunofluorescence confirms co-expressing IT cells in metastatic and primary PDAC

To compare to bulk RNA-seq studies, we clustered pseudo-bulk averages of the malignant cells from each biopsy and observed separation of tumors into those that expressed predominantly basal, classical, or IT signatures (**Supplemental Figure S3.3G-I**). However, individual tumors exhibited significant heterogeneity at the cellular level, with mixing of malignant cell populations expressing at least two and frequently all three cell states within the same patient specimen (**Supplemental Figure S3.3J**). To validate the extensive heterogeneity and the presence of co-expressing IT cells in our metastatic cohort, we used a subtype-specific multiplex immunofluorescence (mIF) panel to categorize single tumor cells by their patterns of marker detection in 10 matched cases from our single cell study (**Supplemental Figure S3.4A**; **Supplemental Table S3.4**; **Methods**). We observed extensive overlap of basal and classical markers within single cells at the protein level, corroborating the existence of co-expressing IT cells using an orthogonal method (**Figure 3.2F**; **Supplemental Figure S3.4B**). Encouragingly, we observed significant correlation within subtype

(average r = 0.52) as compared to between subtypes (average r = 0.06, $P < 10^{-7}$, Student's T test) using this orthogonal method. We also observed high concordance between the two methods, giving confidence that we are accurately sampling the distribution of states present in each sample (average r = 0.45; **Supplemental Figure S3.4C**, white dots). As with scRNA-seq, we observed mixing of basal, classical, and IT cells within individual patient specimens by mIF subtyping. The frequency of co-expressing cells was correlated with balanced representation of pure basal and classical phenotypes within individual samples, consistent with the co-expressing IT phenotype as a transitional state (**Figure 3.2G**). Indeed, none of the tumors evaluated with mIF contained a mix of pure basal and classical phenotypes in the absence of co-expressing IT cells. We also identified co-expressing cells in primary tumor samples which suggests that IT phenotypes may be a general feature of PDAC tumors in both the localized and metastatic settings (**Figure 3.2H**; **Supplemental Figure S3.4D**).

3.3.4 Microenvironment is dominant to KRAS amplifications in determining transcriptional subtype

We next searched for potential molecular regulators of the observed tumor cell transcriptional heterogeneity. In PDAC, the vast majority of tumors harbor clonal *KRAS* point mutations, including all of the PDAC samples in our cohort (**Figure 3.1B**). While point mutations in *KRAS* amplifications in *KRAS* associate with more basal features[19,28], while amplifications of lineage do not appear to determine transcriptional subtype, several studies have suggested that transcription factors like *GATA6* associate with classical phenotypes[19]. To assess for such genotype-phenotype relationships in our single-cell cohort, we inferred copy number variation for common PDAC alterations (*KRAS*, *TP53*, *SMAD4*, and *CDKN2A*) and lineage-associated transcription factors (*HNF4G* and *GATA6*) from scRNA-seq expression data using a previously established Hidden Markov model workflow (**Methods**)[3,13,43]. Encouragingly, we observed a significant association between single-cell inferred *KRAS* copy number gain and basal phenotypes (*P<0.03* Fisher's exact test), and also between inferred *CDKN2A* copy loss and IT phenotypes (*P<0.003* Fisher's exact test; **Figure 3.3A**). While we found that cells derived from samples with inferred *KRAS*

amplifications had a strong preference for the basal subtype, these cells could still span all three phenotypic categories or be predominantly classical within an individual tumor (**Figure 3.3B-D**).

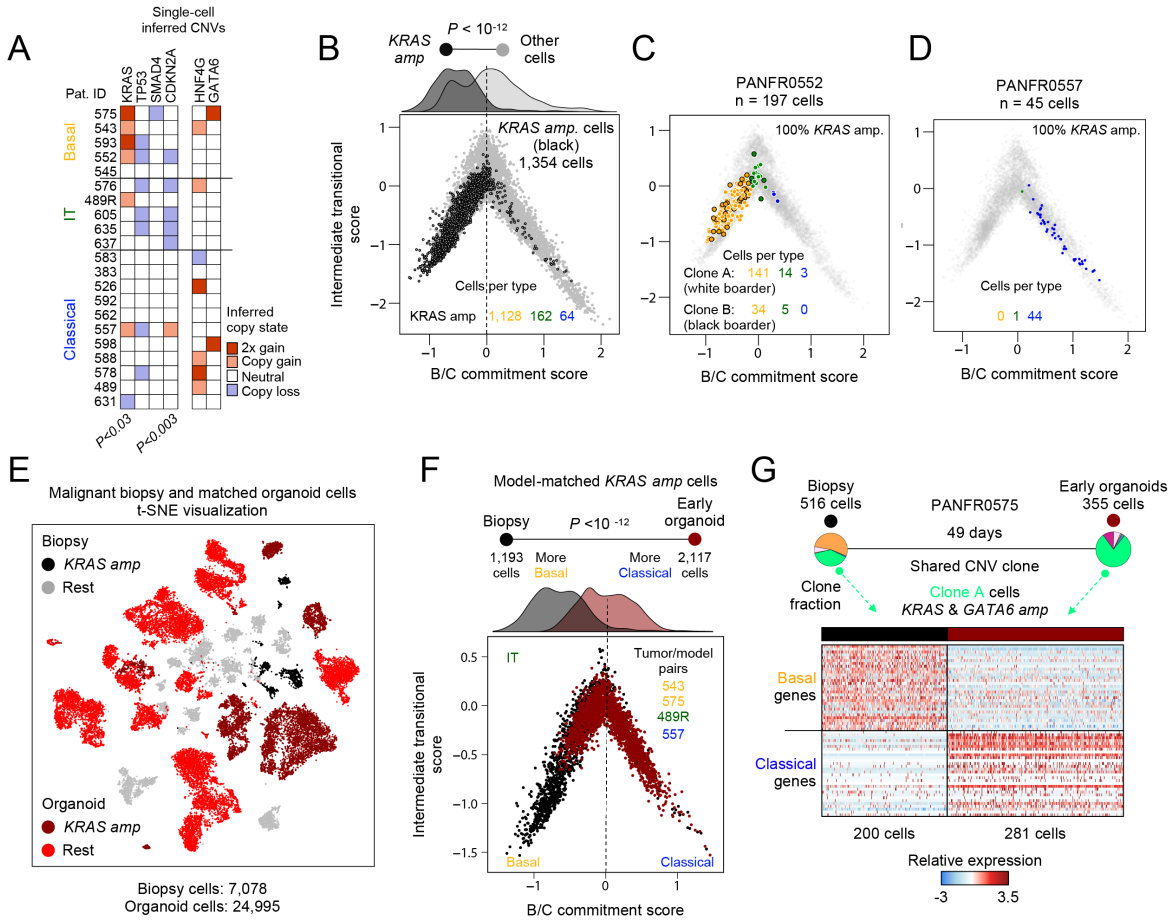


Figure 3.3: Figure 3. Microenvironment dictates phenotype in *KRAS*-amplified tumor cells.

(A) Single-cell inferred copy number alterations for each sample in the biopsy cohort (**Methods**). Tumors are grouped by expression of their dominant subtype based on the clustering in **Supp. Fig. S2.3G**, *P*-values comparing presence of each alteration among the groups (Basal, Classical, IT) are determined by Fisher's exact test. **(B)** Malignant cell state diagram as in **Figure 2.2E** but highlighting all *in vivo KRAS*-amplified tumor cells (black border) across the states. **(C)** Similar to **B**, but highlighting PANFR0552 *KRAS*-amplified malignant cell heterogeneity. White and black borders correspond to separate CNV sub-clones (both *KRAS* amplified) and color fill denotes transcriptional subtype. **(D)** Similar to **B**, but highlighting PANFR0557 *KRAS*-amplified malignant cell heterogeneity. Color fill denotes transcriptional subtype. **(E)** *t*-SNE visualization of all biopsy (grey and black) and matched organoid cells (red and dark red) from iterative passages. *KRAS*-amplified tumor cells from *in vivo* specimens (black) and organoid models (dark red) are highlighted with distinct colors. **(F)** Cell state diagram for all cells with inferred *KRAS* amplifications in biopsy (grey) and organoid (red) microenvironments. *P*-value compares biopsy versus early passage organoid score distributions (top density) and was determined by student's T test. **(G)** Clonal fractions (pie charts) from the *KRAS*-amplified PANFR0575 sample in biopsy and organoid conditions. Heatmap shows the relative expression in single cells from plastic clone A (bright green) in both conditions.

To further examine this genotype-to-phenotype association, we tested if *KRAS* amplification was sufficient to specify the basal phenotype in an *ex vivo* environment. We initiated patient-derived organoid cultures from matched PDAC biopsies and serially sampled them over time with scRNA-seq (**Figure 3.1A**). CNV-confirmed "early" organoid cells (first passage measured, n=2,117 cells) derived from *KRAS*-amplified biopsies maintained this genetic alteration in culture (**Figure 3.3E**, dark red). Despite their genetic stability, cells with inferred *KRAS* amplifications exhibited a profound phenotypic shift from basal *in vivo* to classical *ex vivo* (**Figure 3.3F**). Although selection of specific clones could play a role in this process, most of these models maintained high CNV similarity to their parent tumor at the early time point. For example, we observed that a CNV-defined clone from PANFR0575 with both *KRAS* and *GATA6* amplifications was plastic and shifted from strongly basal *in vivo* to classical in early organoid culture (**Figure 3.3G**). These observations provide strong evidence that phenotypic plasticity is an inherent feature of malignant PDAC cells and demonstrate that *KRAS* amplification alone is not sufficient to lock the basal state. Furthermore, they suggest that the tumor microenvironment can influence phenotype independent of genotype in this context.

3.3.5 Transcriptional heterogeneity is shaped by the microenvironment

Given this strong phenotypic shift even for genetically similar samples, we next examined how *ex vivo* transcriptional phenotypes differed across our larger organoid cohort relative to their cognate patient samples. Globally, unbiased comparison of all malignant biopsy (7,078 cells) and organoid cells (n=14 models, 24,789 cells) revealed unique clusters for each sample and only two clusters that were admixed by donor. These admixed cells exhibited expression programs consistent with non-malignant stromal cells, had low overall CNV scores, and dissipated by later passages (**Supplemental Figure S3.5A-D**; **Methods**). Overall, samples with high tumor-averaged basal or IT phenotypes exhibited lower rates of long-term organoid propagation beyond passage 2 than models derived from classical tumors, where the majority established long-term cultures (**Figure 3.4A**). When comparing early passage CNV-confirmed organoids to their cognate patient tissue, culture in an *ex vivo* microenvironment caused greater deviation in transcriptional phenotype than CNV-defined genotype (**Figure 3.4A**, $P < 10^{-6}$ Student's T test; **Methods**).

We next assessed which specific tumor cell attributes contributed to phenotypic divergence in the *ex vivo* microenvironment. As with the *KRAS*-amplified samples, we observed a striking

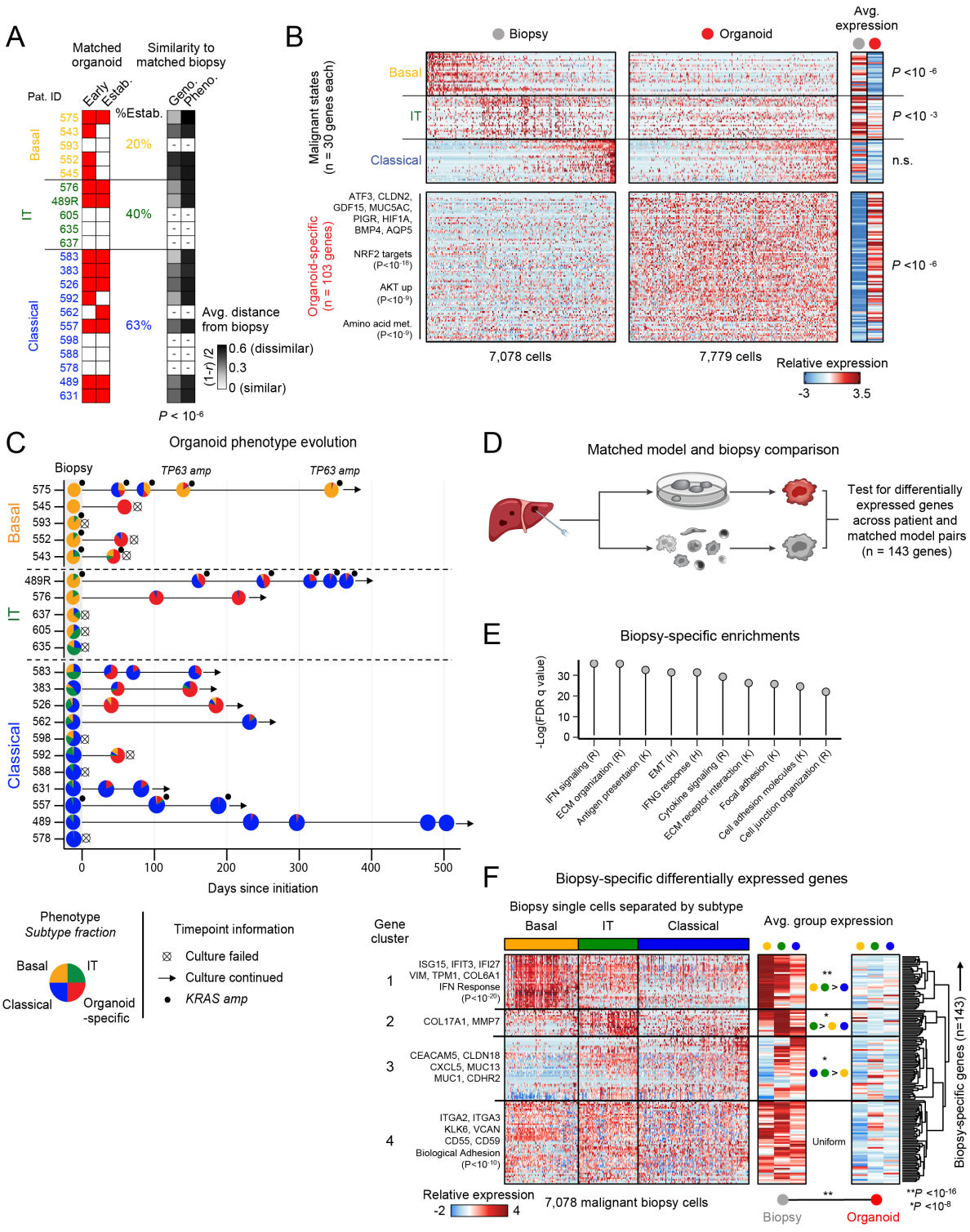


Figure 3.4: Organoid culture microenvironment selects against the basal state with phenotypic evolution over time. **(A)** Sampling from each model initiated as an organoid. Red fill represents measurements at an Early time point and if that biopsy established a long-term culture (Estab.). Right grey scale heat indicates the distance (**Methods**) between each biopsy-early organoid pair for CNVs (Geno.) or transcriptional subtype (Pheno.). *P*-value for Geno. vs Pheno. differences determined by student's T test. **(B)** Relative expression for the malignant programs (top) and organoid-specific genes (bottom) in biopsy cells (left) and their matched, early passage organoid cells (n=13 models; right). Parenthetical *P*-values (left) indicate hypergeometric test for enrichment of pathways in the indicated gene clusters. Far right heat is average expression for all genes in each group, *P*-values determined by student's T test. **(C)** Swimmer's plot shows the evolution of organoid phenotype in the culture microenvironment. Each point indicates a passage when organoids were sampled with scRNA-seq, and pie chart fill indicates the fraction of single cells binned as each transcriptional subtype. **(D)** Schematic for matched tumor-organoid differential expression analysis. **(E)** Top differentially expressed genes *in vivo* (143 genes) are TME-associated and enrich for TME-associated pathways. All top enrichments shown are highly significant (*P*-value < 10^{-12}). **(F)** Hierarchical clustering in biopsy cells (columns) of the relative expression for the 143 TME-associated genes preferentially expressed *in vivo* (rows). Cells are binned in the single-cell heatmap and the averages at right by their originating tumor's average transcriptional subtype. Gene-level averages are split by biopsy (left) and organoid cells (right). Parenthetical *P*-values (left) indicate hypergeometric test for enrichment of pathways in the indicated gene clusters. For within-group differences in expression for biopsy averages, *P*-values are computed by one-way ANOVA followed by Tukey's HSD and compare averaged expression of each gene cluster between cells from different biopsy subsets (middle heatmap; **P*-value < 10^{-8}; ***P*-value < 10^{-16}). Overall biopsy versus organoid average expression difference for all 143 genes is determined by Student's T test.

decrease in basal gene expression ($P < 0.000001$) and, to a lesser but still significant extent ($P < 0.001$), the IT program (**Figure 3.4B**, **top**). By contrast, aggregate classical gene expression remained largely unchanged in organoid conditions (**Figure 3.4B**, **top**). This loss of basal expression was surprising, given the more clinically aggressive and proliferative nature of basal tumors *in vivo* (**Supplemental Figure S3.3A**)[21,22]. Organoid-specific gene expression features that were not present *in vivo* also emerged, including markers of epithelial identity, oxidative stress response pathways (e.g., NRF2 target genes), and amino acid metabolism (hereafter collectively referred to as "organoid-specific" gene expression; **Figure 3.4B**, **bottom**; **Supplemental Table S3.5**). In general, models assumed a more classical or organoid-specific phenotype over time in culture regardless of their parent tumor's transcriptional identity (**Figure 3.4C**). Most models derived from basal or IT tumors exhibited early phenotypic deviation and cessation of growth within 100 days of initiation (e.g., PANFR0552; **Supplemental Figure S3.5E**) or outgrowth of only a sub-clone in culture (e.g., PANFR0489R; **Supplemental Figure S3.5F**). Classical tumors,

meanwhile, tended to maintain their genotype and phenotype both early in culture and at later passages (e.g., PANFR0631; **Figure 3.4C**; **Supplemental Figure S3.5G**, clone A).

To better understand the contribution of clonal selection to this process, we performed linked genotype and phenotype assessment from iterative passages. We identified CNV-defined subclones in the parental biopsy and its associated serial organoid samples, and then assessed how the distribution of transcriptional states within each subclonal population evolved over time in culture (**Methods**). In both PANFR0489R and PANFR0575, *KRAS* amplification status remained invariant over time, but we observed significant phenotypic plasticity and clonal selection in both cases. In PANFR0489R, the predominantly basal clones *in vivo* rapidly decreased in abundance while other rarer clones with classical or organoid-specific phenotypes emerged as the dominant ones (**Supplemental Figure S3.5H**). In contrast, *in vivo* dominant clones from PANFR0575 were largely maintained at early passages but diverged significantly in their phenotype, transiently expressing more classical and organoid-specific phenotypes at passages 2 and 3 before eventually regaining basal transcriptional expression after >100 days in culture (**Supplemental Figure S3.5I**). Notably, the clones that came to dominate in PANFR0575 organoid culture (clones D and E, **Supplemental Figure S3.5I**) carried inferred *TP63* amplifications, a squamous-specifying transcription factor[44], suggesting that certain genotypes, though rare, may still exert a strong effect despite opposing signals from the microenvironment. Taken together, these findings emphasize the importance of optimizing culture conditions and performing deep molecular characterization of patient-derived model systems to ensure faithful representation of the tumor.

Divergence from *in vivo* phenotype, despite relative similarity in genotype, suggested that the TME has a strong influence in determining PDAC cellular state. For each biopsy-organoid pair, we used differential expression to nominate transcriptional programs that were present *in vivo* but missing from *ex vivo* culture (**Figure 3.4D**; **Methods**). Broadly, genes preferentially expressed by malignant cells *in vivo* were related to soluble cytokine signaling, cell-cell communication, and tumor-microenvironment interactions, highlighting the absence of this crosstalk in organoid culture (**Figure 3.4E**). Hierarchical clustering revealed subtype-dependent expression patterns for these *in vivo*-specific genes (**Figure 3.4F**; **Supplemental Table S3.5**). For example, interferon response and EMT genes were significantly upregulated in basal and IT malignant cells *in vivo* (clusters 1 and 2, **Figure 3.4F**), while genes associated with cell-cell interactions and surface glycoproteins were more strongly expressed in IT and classical cells (cluster 3, **Figure 3.4F**).

Genes related to biological adhesion were more uniform in their expression across the subtypes (cluster 4, **Figure 3.4F**). The relative absence of these TME-crosstalk genes in organoid culture and their differences in expression across transcriptional subtypes *in vivo* suggest that TME signals may play a role in specifying tumor cell phenotypes.

3.3.6 Non-malignant composition of the metastatic microenvironment

The presence of TME-associated expression patterns in cancer cells *in vivo* suggested there may be subtype-dependent structure to, and instructive signaling from, the metastatic TME; however, relatively little is known about the structure and composition of the metastatic microenvironment in PDAC. We first analyzed the non-malignant cells (n=14,811) in the metastatic niche to further subclassify cell types and provide a more complete picture of the immune/stromal composition of metastatic disease (**Figure 3.5A**). Sub-clustering of T/NK cells revealed 4 cell types—*CD4*+ T, *CD8*+ T, NK, and *CD16*+ (*FCGR3A*+) NK cells—each expressing the corresponding established markers (**Supplemental Figure S3.6A**,**B**; **Methods**). Similarly, an unsupervised examination within the monocyte/macrophage compartment revealed a tripartite continuum for tumor associated macrophages (TAMs), similar to one recently described in colorectal cancer, comprised of inflammatory *FCN1*+ "monocyte-like" TAMs, *C1QC*+ phagocytic TAMs, and *SPP1*+ angiogenesis-associated TAMs (**Supplemental Figure S3.6C**,**D**; **Supplemental Table S3.2**)[45,46]. Representative marker expression across all previously described non-malignant cells is summarized in **Supplemental Figure S3.6E**.

Although most samples in our cohort were taken from liver metastases (19/23), several originated from other sites including the omentum, adrenal gland, and peritoneum (**Figure 3.1B**, **"**other"). Interestingly, while we found equal distribution of immune cells among the anatomical sites, mesenchymal cell populations clustered predominantly by the location of the metastatic lesion (**Figure 3.5B**,**C**). Excluding adrenal-specific endocrine cells (**Figure 3.4C**; subset 4, 40 cells), we identified 3 mesenchymal subclusters with relatively uniform expression for canonical cancer-associated fibroblast (CAF) markers (**Figure 3.5C**; **Supplemental Figure S3.6F**). PCA of these cells revealed a continuum of states along PC2, with uniform expression of the previously described myofibroblast (myCAF) signature[34,36] but further separating into cells favoring high expression of dermal fibroblast-like genes (PC2 low, *FAP, PRXX1, SFRP2*) or pericyte-like genes (PC2 high, *RGS5, MCAM, TBX2*; **Supplemental Figure S3.6G-I**; **Supplemental Table S3.2**)[47-

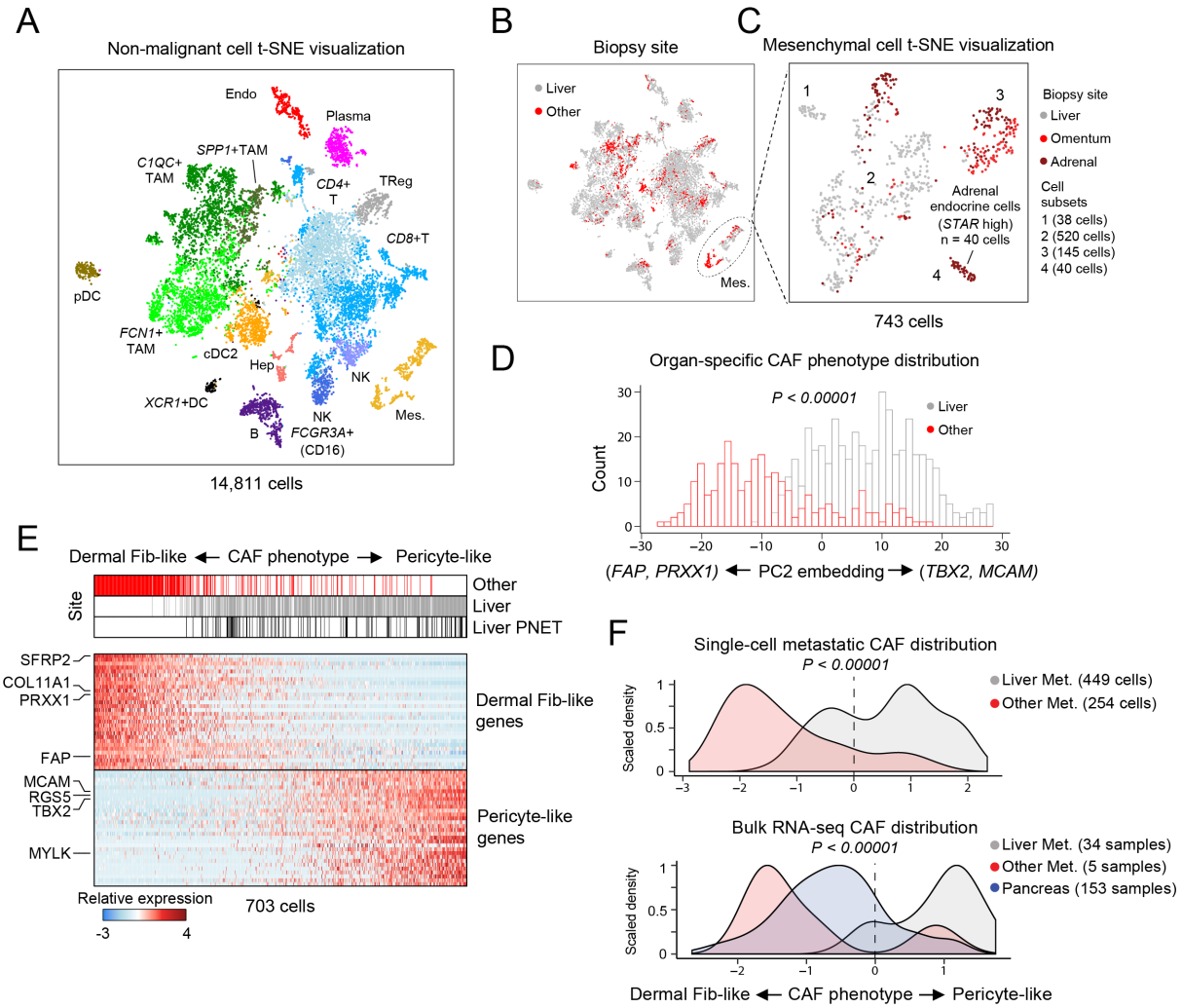


Figure 3.5. Immune heterogeneity and distinct fibroblast phenotypes exist in the liver metastatic microenvironment. **(A)** *t*-SNE visualization of non-malignant cells identified in the metastatic microenvironment, abbreviations are the same as in **Figure 3.1D** (TAM, tumor associated macrophage; NK, natural killer). **(B)** Same visualization as in **A**, but cells are colored by sampling site (Liver, grey; Other, red). Only the mesenchymal cells (dotted circle, Mes.) have appreciable separation by anatomical site. **(C)** *t*-SNE visualization of sub-clustering (SNN) performed on mesenchymal cells colored by their anatomical site. Cell subsets (1-4) determined by SNN clustering. **(D)** Frequency of CAFs (y axis, cell count) across PC2 scores, colored by site of biopsy tissue. *P*-value determined by student's T test. **(E)** Heatmap for relative expression of the Dermal Fibroblast-like (PC2 low) and Pericyte-like (PC2 high) programs. Anatomical site is shown for each cell (top). **(F)** Density plots for CAF phenotype score in single cells from our metastatic cohort (top) or previously published PDAC bulk RNA-seq profiles (bottom)[15,18], fill indicates anatomical site. *P*-value determined by student's T test (top) or by ANOVA followed by Tukey's HSD (bottom).

[52]. PC3 described a small subset of cells, derived largely from a single tumor (PANFR0489R), that were highly consistent with the previously established inflammatory fibroblast (iCAF) program (**Supplemental Figure S3.6H-J**)[34,36].

While tumors from each location contained both mesenchymal subsets, we noted a strong organ-specific skewing along PC2 with pericyte-like phenotypes being preferentially associated with liver biopsies (**Figure 3.5D**,**E**; **Supplemental Figure S3.6K**). To validate these observations in larger cohorts, we assessed bulk RNA-seq datasets using these dermal fibroblast- and pericyte-like CAF signature scores and observed a similar predilection for the pericyte-like expression program in liver metastases (**Figure 3.5F**; **Methods**). Interestingly, tumors in the pancreas (n = 153 samples) favored expression of the dermal fibroblast-like program, suggesting a substantially different mesenchymal microenvironment in primary versus liver metastatic PDAC (**Figure 3.5F**). Thus, we observed diverse immune and stromal cell types in the metastatic TME and identified site specific mesenchymal features unique to the liver metastatic niche compared with primary disease.

3.3.7 Transcriptional subtypes associate with distinct immune microenvironments

After cataloging the cell types in the metastatic TME, we searched for associations between malignant subtype and the immune microenvironment. For each tumor sample, we first computed the fractional representation of each non-malignant cell type per biopsy. Five tumors were excluded from this analysis based on low cell counts (<200 cells) or indeterminant transcriptional subtype (PanNET or no tumor cells captured; **Supplemental Figure S3.6L**). To describe the overall microenvironmental composition for each tumor, we applied Simpson's diversity index, a measure of biodiversity commonly used in ecology to describe the number of species (cell types) present in an ecosystem (tumor) and their relative abundance. We observed that tumors with more classical or IT phenotypes exhibited greater microenvironmental diversity, while strongly basal tumors had a more homogeneous TME (**Figure 3.6A**). Hierarchical clustering over the relative abundance of each non-malignant subset across the biopsy cohort revealed the specific cell types driving these overall diversity differences (**Figure 3.6B**,**C**). Specifically, C1QC+ TAMs dominated the microenvironments of strongly basal tumors, and both CD8+ and CD4+ T cells were significantly depleted in basal contexts compared to the rest of the samples in the cohort (**Figure 3.6C**,**D**). T cells most often originated from biopsies with higher IT malignant fractions (**Figure 3.6B**,**C**) and their abundance was positively correlated with this malignant phenotype in our cohort (**Figure 3.6E**). We also broadly observed these patterns within TCGA bulk RNA-sequencing data of other epithelial malignancies[53], where we observed evidence for reduced levels

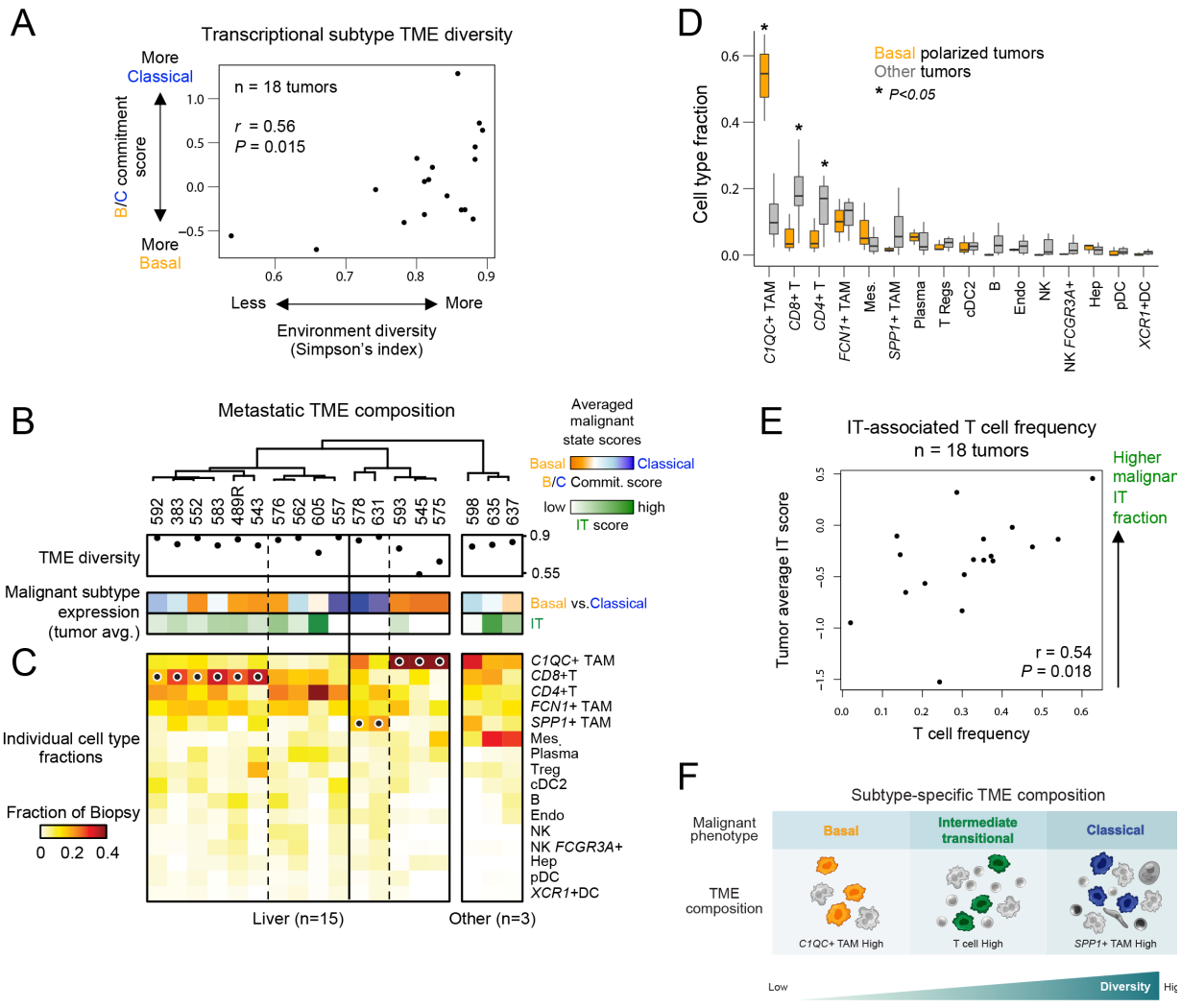


Figure 3.6: Transcriptional subtypes associate with distinct metastatic microenvironments.

(A) Correlation between microenvironment diversity (Simpson's Index, x axis) and the average malignant basal-classical commitment score for each biopsy (y axis). **(B)** Dot plot indicates the Simpson's Index calculated for each biopsy and heat bars indicate each tumor's average malignant cell expression for each of the malignant transcriptional programs. **(C)** Fraction of each non-malignant cell type (heat, rows) in each biopsy sample (columns). Dots indicate top statistically significant cell type frequency differences calculated using Kruskal-Wallis test with multiple hypothesis correction. Samples are ordered as in **B. (D)** Box plots compare cell type fraction between the basal polarized tumors with low diversity (PANFR0593, 575, 545) and all others. *P*-value determined by student's T test. **(E)** Correlation between T cell fraction and IT malignant score. (**F**) Schematic summarizing associations between microenvironmental diversity, non-malignant infiltrates, and tumor subtype.

of immune-related gene expression in tumors with high basal/squamous gene expression (**Supplemental Figure S3.6M**, cluster 4). Taken together, these findings suggest coordinated interactions between malignant phenotypes and the local TME with decreased immune cell diversity and a greater degree of immune exclusion associated with basal contexts (**Figure 3.6F**).

3.3.8 The soluble microenvironment shapes PDAC cellular phenotypes

Based on our observations that: 1) the microenvironment influences malignant phenotype independent of genotype; 2) gene expression programs associated with cytokine signaling, EMT,

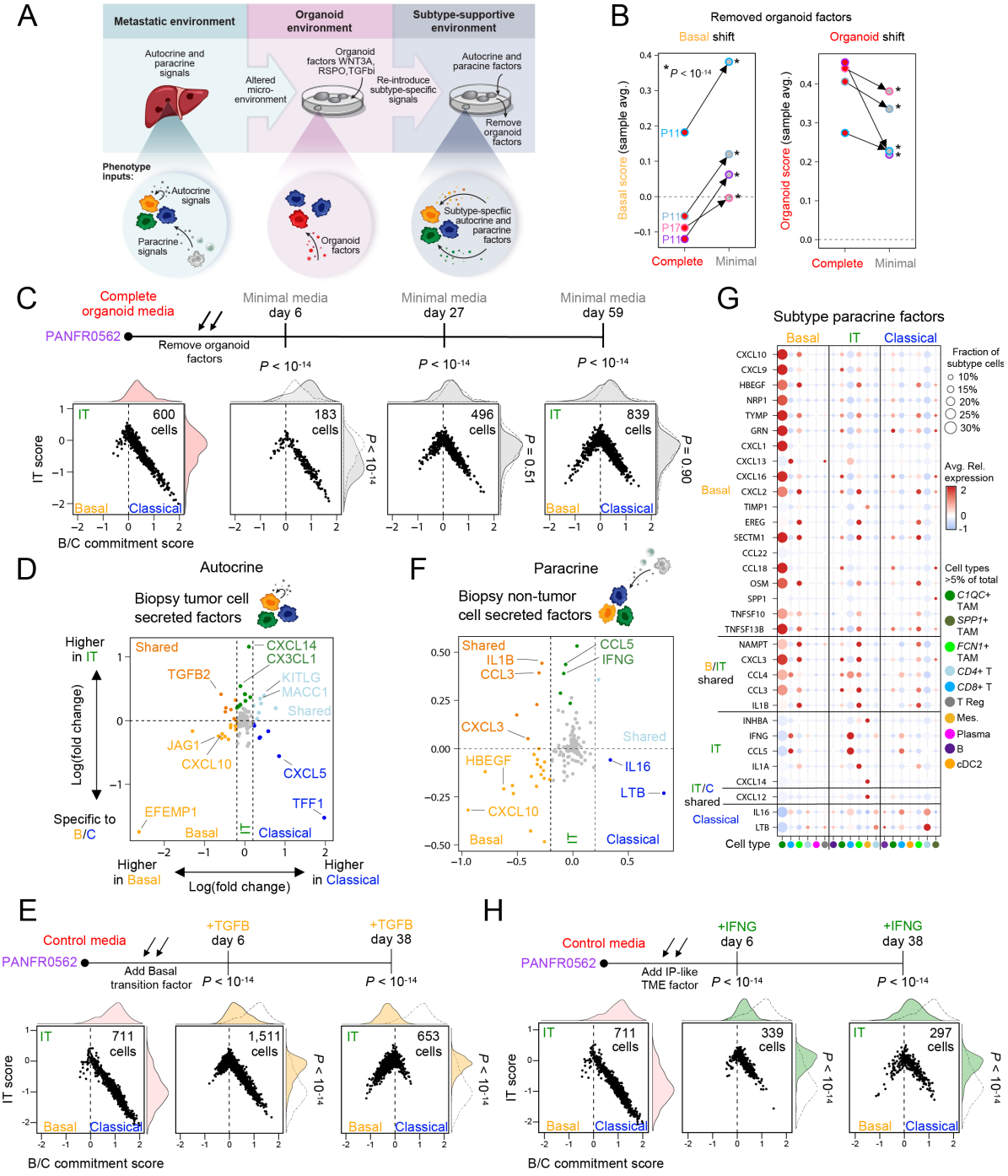


Figure 3.7: Tumor subtype-specific secreted microenvironmental factors rescue malignant transcriptional heterogeneity.

(A) Schematic describing microenvironmental inputs *in vivo* ("Metastatic environment"*)* versus *ex vivo* ("Organoid environment") to tumor phenotype. Right panel ("Subtype-supportive environment") describes an overall strategy to recover malignant transcriptional heterogeneity by removing organoid factors (**B, C**) and adding state-specific autocrine (**D, E**) or paracrine (**F**-**H**) factors. **(B)** Tied dot plot represents the sample average basal score (left) and organoid-specific score (right) in the indicated conditions. Lines tie samples and color outlines indicate sample identity as in **Figure 3.1B**. *P*-value compares respective single cell distributions within models and was calculated by student's T test. **(C)** Cell state diagrams for organoid cells cultured in complete medium or at 3 time points in minimal media. *P*-values for group differences between B/C commitment (top) and IT scores (right) were calculated by ANOVA followed by Tukey's HSD. *P*-values displayed are for that timepoint vs. the complete media condition.**(D)** Differential expression (Wilcoxon rank sum test) for known secreted factors by *in vivo* tumor cells (autocrine) between basal and classical (x axis) and IT malignant cells and the rest (y axis). Subtype-specific genes that pass significance after multiple hypothesis correction ($P < 0.05$) are colored by their group association. **(E)** Cell state diagrams with marginal density plots for organoid cells cultured in control medium (OWRNA, reduced organoid medium) or at 2 time points in control media with TGF-β. *P*-values for group differences between B/C commitment (top) and IT scores (right) were calculated by ANOVA followed by Tukey's HSD. *P*-values displayed are for that timepoint vs. the control media condition. **(F)** Differential expression (Wilcoxon rank sum test) for known secreted factors by all non-malignant cells (paracrine) found in basal and classical (x axis) and IT biopsies and the rest (y axis). Subtype-specific genes expressed by non-malignant cells that pass significance after multiple hypothesis correction ($P < 0.05$) are colored by their group association. **(G)** Dot plot for the subtype-specific significant differentially expressed paracrine factors. Subtype-specific non-malignant cell types (columns) and significant genes (rows) are binned by subtype association as in **Figure 3.6C** and **Figure 3.7F**. Dot size represents that cell type's fraction within tumors of each subtype, and fill color indicates average expression. Only cell types with a fractional representation >5% from each subtype are visualized. **(H)** Cell state diagrams with marginal density plots for organoid cells cultured in control medium (OWRNA, reduced organoid medium, as in **E**) or at 2 time points in control media with IFNγ. *P*-values for group differences between B/C commitment (top) and IT scores (right) were calculated by ANOVA followed by Tukey's HSD. *P*-values displayed are for that timepoint vs. the control media condition.

and cell-cell interaction are enriched *in vivo* but missing from cells cultured as organoids; and, 3) malignant states and immune cell infiltration are coordinated in a subtype-specific manner, we hypothesized that incorporation of soluble factors specific to the TME of each transcriptional subtype may drive tumor cell state shifts (**Figure 3.7A**). Complete PDAC organoid media (**Supplemental Table S3.6**)[25,40] contains various growth factors that could skew malignant transcriptional state, so we first tested the effects of withdrawing various soluble factors. We cultured four organoid models in media without any additives ("Minimal" media, containing only Glutamax, anti-microbials, HEPES buffer, and Advanced DMEM/F12 media; **Figure 3.7B**;

Supplemental Table S3.6; **Methods**). We observed a robust increase in basal gene expression and a decrease in organoid-specific gene expression in specimens cultured for 6 days in minimal media relative to those in complete organoid media ("Complete", **Figure 3.7B**). Although we found that the fraction of cycling cells in minimal media decreased, the organoids continued to grow under these conditions and exhibited stable CNV profiles, indicating that these responses were unlikely to be driven by acute selection (**Supplemental Figure S3.7A**,**B**). We cultured one model, PANFR0562, in minimal media for a longer duration and observed that the phenotypic distribution shifted even further toward IT and basal phenotypes (**Figure 3.7C**), demonstrating that recovery of all three states is possible *ex vivo*. Since minimal medium lacks both serum and mitogens to support prolonged cell growth, we also tested whether culturing organoids in a reduced organoid media formulation ("OWRNA", complete organoid media with removal of WNT3A, RSPONDIN-1, NOGGIN, and A83-01; **Supplemental Table S3.6**; **Methods**) supported proliferation while allowing expression of basal and IT phenotypes. We found that organoids maintained under OWRNA conditions began to express basal and IT features while also strengthening classical gene expression and continuing to proliferate (**Supplemental Figure S3.7C**).

To assess whether these microenvironment-driven effects on transcriptional states were specific to organoid models or also observed in other cell culture models, we examined PDAC cell lines, as these are also commonly used to study PDAC biology but are grown in different culture conditions. We compared bulk RNA expression data from patient tumors (n=219)[15,18], our own organoid cohort (n=44), and established cell lines (n=49, CCLE)[54,55] and observed strong culture method-dependent phenotypic skews wherein most organoid models expressed classical phenotypes while cell lines exhibited basal phenotypes (**Supplemental Figure S3.7D**,**E**). This observation suggests neither platform accurately represents the full repertoire of transcriptional states seen in patients and provides additional evidence that environmental conditions can profoundly influence transcriptional state. We ruled out the effects of extracellular matrix dimensionality from media formulation by culturing established 3-dimensional (3D) organoid models as 2-dimensional (2D) cell lines on tissue culture plastic in the same organoid media—this had little effect on transcriptional subtype across the models tested (**Supplemental Figure S3.7F**). Next, we took each model type (cell lines and organoids) and cultured it in the reciprocal media condition to ask whether media alone could influence transcriptional subtype. Organoid cells

grown in standard cancer cell line medium ("RP10", RPMI-1640 with 10% fetal bovine serum) gained expression of basal programs (**Supplemental Figure S3.7C**), while CFPAC1 (an established PDAC cell line) lost basal and classical features and gained organoid-specific gene expression when grown in complete organoid media ("Complete media", **Supplemental Figure S3.7G**). Taken together, these findings demonstrate that the microenvironment is an instrumental contributor to shaping malignant phenotypes in PDAC. Moreover, the cell state plasticity suggests the possibility of testing subtype-specific conditions to support the full repertoire of *in vivo* phenotypes.

3.3.9 Applying subtype-specific TME signals drives patient-relevant subtype heterogeneity

Finally, we hypothesized that specific factors from subtype-specific TMEs could recover clinically relevant transcriptional heterogeneity *ex vivo* (**Figure 3.7A**). *In vivo,* the secreted factor milieu surrounding tumor cells originates from at least two sources that may influence malignant phenotype: tumor cells themselves ("autocrine" factors) and non-tumor cells ("paracrine" factors, **Figure 3.7A**). First, to nominate possible autocrine signals, we identified tumor cell secreted factors specific to the three subtypes and noted distinct cytokines expressed by each (**Figure 3.7D**; **Supplemental Table S3.7**; **Methods**). Since malignant cells derived from predominantly basal and IT tumors lose their phenotype in organoid culture, we first tested factors specific to IT and basal states *in vivo*. *TGFB2* was the top differentially expressed secreted factor shared by tumor cells in both basal and IT TMEs (**Figure 3.7D**). Organoids cultured with TGF-β ligands exhibited a loss of classical expression programs and a near complete shift toward IT and basal phenotypes, matching what we observed *in vivo* (**Figure 3.7E**). Reemergence of basal phenotypes in both minimal media (**Figure 3.7C**), and TGF-β conditions (**Figure 3.7E**) suggest that different types of microenvironmental pressure can lead to the basal phenotype. Moreover, they suggest that culture conditions can be tuned to achieve compositional differences spanning pure classical, heterogenous, and pure basal phenotypes, akin to those seen *in vivo*.

Using a similar approach, we next searched for differentially expressed paracrine factors supplied by the non-tumor cells in the TME from each subtype. Here, we noted an increasing number of differentially expressed factors in IT and basal contexts, likely reflecting the specific immune cell type enrichments: TAM and T cell dominant in basal and IT TMEs, respectively (**Figure 3.6A-C**; **Figure 3.7F**; **Supplemental Table S3.7**). We then mapped each subtype-specific

paracrine factor to its cognate cell type to summarize the overall cell type and secreted factor combinations that shape the subtype-specific TMEs in metastatic PDAC (**Figure 3.7G**). Interestingly, we found that *IFNG* originating from *CD8+* T cells was most highly expressed in the IT TME (**Figure 3.7F**,**G**). This was consistent with a relatively higher T cell fraction in IT tumors (**Figure 3.6B**,**F**) and the relative increase in IFN responsive gene expression in IT and basal tumor cells (**Figure 3.4E**,**F**). Given these corroborating correlative data, we directly tested whether exogenous IFNγ could induce transcriptional plasticity towards an IT state. Cells exposed to IFNγ showed a dramatic shift toward the IT state with concomitant decrease in expression of classical signatures (**Figure 3.7H**). In contrast with exogenous TGF-β (**Figure 3.7E**), microenvironmental IFNγ seemed to more specifically induce an IT state, as these cells did not fully transition to basal phenotypes at later timepoints (**Figure 3.7H**). These findings demonstrate that the microenvironment plays a critical role in specifying tumor transcriptional phenotypes and provide evidence for significant PDAC tumor cell plasticity in response to microenvironmental cues.

3.4 Discussion

Here, by linking single-cell profiling of *in vivo* patient specimens to matched organoid models, we have built an essential comparative dataset to disentangle the contributions of cell-intrinsic versus -extrinsic factors to cancer cell transcriptional states in metastatic PDAC. We leveraged the precision afforded by scRNA-seq to identify a new PDAC cell state that co-expresses the basal and classical programs and behaves as a transitional intermediate between the basal and classical subtypes. Importantly, the identification of large fractions of co-expressing IT cells in human tumor biopsies using both mIF and scRNA-seq suggests interconversion between the classical and basal subtypes occurs frequently in response to various cues *in vivo* and implies that this intermediate state may be a hallmark of intratumoral plasticity and tumor cell transcriptional evolution. In fact, in contrast to prior reports[19], all tumors that had mixed but discrete populations of basal and classical cells also exhibited proportional fractions of co-expressing IT cells. Our matched organoid studies provide strong evidence that this extensive transcriptional heterogeneity is heavily influenced by the microenvironment, a finding that is further reinforced by the identification of subtype-dependent TME structure. As such, this work provides a detailed description of the PDAC metastatic niche, critical insight into the role of the microenvironment in

determining cancer cell phenotype in PDAC, and a general framework for discovering and manipulating these relationships across cancer contexts.

Although mutations in *KRAS* play a critical role in pancreatic oncogenesis, PDAC cells have also been shown to adopt more RAS-independent phenotypes as a mechanism of resistance to KRAS suppression[29]. Our findings help to reconcile these opposing observations by suggesting that KRAS target gene expression is more strongly associated with the IT cell state than either basal or classical extremes. This finding suggests that while upregulation of KRAS signaling by amplification or other mechanisms may play an important role in the transition toward the basal state[19,28], it may become less functionally important once this state transition is complete. Furthermore, the presence of IT cells enriched for *KRAS* and inflammatory response gene expression is reminiscent of phenotypes seen in mouse models that suggest inflamed progenitor-like cells as those that tolerate *KRAS* mutations and initiate tumorigenesis[56,57].

Our single-cell data support the association between *KRAS* amplifications and the basal state *in vivo*; however, when we compared our matched *KRAS* amplified biopsy and organoid cells, we saw that this genotype did not lock cells into the basal state, and that microenvironmental conditions were a dominant factor in determining tumor cell transcriptional subtype. Serial sampling of organoid models across successive passages demonstrated both phenotypic drift and sub-clonal outgrowth, mirroring the genetic evolution of PDXs and cell lines in culture[58,59], and highlighting the complex interplay between genetics and microenvironmental influences on transcriptional plasticity and clonal selection. This facile transition between subtypes has important implications for drug treatment, and future studies using lineage tracing approaches are needed to better understand the evolutionary dynamics in this system and how to track and exploit these processes therapeutically. Additional studies into the epigenetic regulatory mechanisms underlying PDAC state transitions will also be a critical next step in further delineating the relationships between genotype, microenvironment, and phenotype.

Although we have identified co-expressing IT cells in both primary and metastatic tumors, the transcriptional programs associated with co-expression may differ between these contexts. We hypothesize that the basal state may be a common phenotypic endpoint for PDAC tumor cells in response to microenvironmental stress, with superimposed transcriptional variation depending upon the specific stressors a given tumor cell must overcome to reach this state. Supporting this concept is the observation that cells exhibiting basal phenotypes show concomitant expression of

EMT, IFN response, or hypoxia response signatures, and these expression patterns may be driven by the specific microenvironment[29,60]. In addition, our finding that diverse microenvironmental signals, including nutrient deprivation ("Minimal media"), autocrine and stromal signals (TGF-β), and immune signals (IFNγ), induce the transition away from the classical subtype further supports this conclusion. We postulate that IT intermediates likely house similar context-dependent complexity depending on the tissue of residence.

Similar to malignant cells, the non-malignant cell types in the metastatic TME were varied in phenotype and overall composition. Although we mainly sampled liver metastases, we identified strong differences between mesenchymal populations from different biopsy sites. We observed that the liver metastatic niche was enriched for pericyte-like myofibroblasts[48-51], while other sites of metastasis and primary disease were enriched for dermal fibroblast-like phenotypes. Given the pivotal role that has been suggested for the fibrotic TME in primary disease[37,61], these findings carry important implications for targeting the stromal compartment in primary versus metastatic PDAC. For example, inhibitors targeting FAP have recently shown preclinical efficacy[62], but we observe FAP expression favors dermal fibroblast-like cells but not pericyte-like myofibroblasts which are more prevalent in liver metastases. As such, examination of these fibroblast phenotypes across larger sample sets may help to identify additional clinically relevant variation in tumor-fibroblast crosstalk, and site-specific combinatorial strategies may be needed to effectively target the PDAC tumor stroma.

We show how scRNA-seq can be employed to define the structure of the metastatic niche and uncover formerly unappreciated relationships between tumor transcriptional phenotype and the local TME. Although traditionally thought of as a uniformly "immune-cold" tumor, our findings highlight that the immune microenvironment in metastatic PDAC harbors a layer of complexity closely linked to tumor cell transcriptional subtype that may provide new avenues for therapeutic targeting. Notably, we observed high levels of *IFNG* expression by CD8+ T cells and coordinated elevation in IFN response gene expression in IT and basal malignant cells. We recapitulated this shift from a classical to a more IT state in organoid models exposed to IFNγ, suggesting that malignant adaptation to signals from the TME may contribute to driving IT and basal phenotypes. Similar to the relationship between inflammation and tumorigenesis[56,57], we speculate that as tumors become inflamed and immune-activated, malignant cells display enhanced plasticity, transition to an IT state in response, and then progress to a fully basal phenotype with

concomitant immune evasion and exclusion. These relationships may have implications for PDAC response to immunotherapy given that a productive immune response may promote more aggressive basal phenotypes[57,60]. Notably, we observed evidence for basal expression signatures with a corresponding paucity of immune cell type signatures in multiple other epithelial cancers, suggesting that coordination of malignant and immune responses in basal contexts may be a broadly relevant phenomenon across many cancer types. Additional studies with co-culture, mouse models, or serial samples from patients on active immunotherapy may further clarify these coordinated and reciprocal tumor-immune interactions.

More generally, our approach using matched *in vivo* malignant populations as a reference for *ex vivo* perturbations and model generation provides a critical framework for understanding the signals that drive clinically relevant phenotypes but are missing from organoid and cell line cultures. The genetic evolution of *ex vivo* models is a well-established phenomenon which carries functional consequences[59,63]. Our study highlights similar *ex vivo* evolution for transcriptional variation, but also provides a strategy to rescue malignant phenotypes by re-introduction of soluble signals needed for their support *in vivo*. This approach may offer a more tractable system for state-specific high throughput screening compared with more complex heterotypic co-cultures or PDX systems. With a catalogue of matched *in vivo* phenotypes as a reference, this workflow empowers not only model fidelity, but enhances our ability to learn the phenotypic boundary conditions for individual tumors. For example, we can begin to define whether certain pressures induce cell state transitions in specific subsets of *ex vivo* models and identify which combinations of factors impede or synergistically enhance these transitions. Furthermore, these studies highlight how model generation in different growth contexts—organoids, cell lines, spheroids—may lead to the identification of emergent tumor cell properties. Learning these rules across different tumor contexts and understanding which non-malignant cell types participate *in vivo* would allow for the full appreciation of the symbiotic relationships within tumor ecosystems and provide a valuable foundation for leveraging microenvironmental manipulation to control tumor cell phenotype and behavior.

In sum, our data demonstrate coordinated phenotypic evolution driven by reciprocal interactions between malignant cells and the TME in PDAC. Just as we consider therapeutic combinations to target tumor cell intrinsic properties, paracrine interactions with the TME may equally drive tumor cell phenotype and thus require consideration in designing combination

strategies. We provide a framework for relating malignant cells, the TME, and patient-derived model systems that may be applicable in other tumor types with clinically relevant transcriptional variation across the malignant and microenvironmental compartments.

3.5 Methods

*3.5.1 Tissue collection and dissociation***.** Investigators obtained written, informed consent from patients with pancreatic cancer for Dana-Farber/Harvard Cancer Center Institutional Review Board (IRB)-approved protocols 11-104, 17-000, 03-189, and/or 14-408 for tissue collection, molecular analysis, and organoid generation. Core needle biopsy specimens were collected and the first core was sent for pathologic analysis. One or more additional cores were then allocated for scRNA-seq and organoid generation.

Samples were minced into small portions using a scalpel and then digested at 37°C for 15 minutes using digest medium that consisted of human complete organoid medium (see below), 1 mg/mL collagenase XI (Sigma Aldrich), 10 µg/mL DNase (Stem Cell Technologies), and 10 µM Y27632 (Selleck)[25]. In our initial process optimization, we found that dissociation times below 30 minutes, while not always completely digesting all biopsy material and potentially affecting the representation of difficult to dissociate cell types (e.g., fibroblasts), resulted in greater cell viability and improved RNA quality downstream. After digestion, cells were washed, treated with ACK lysing buffer (Gibco) to lyse red blood cells, washed again, and counted using a hemocytometer with 0.4% Trypan blue (Gibco) added at 1:1 dilution for viability assessment. We allowed residual tissue chunks to settle before selecting a predominance of single cells for counting and Seq-Well processing. We allocated between 10,000 and 15,000 viable cells per Seq-Well array based upon total cell counts, and where possible we prepared two arrays per sample. Most samples were processed and loaded onto Seq-Well arrays within 2-3 hours of biopsy acquisition.

3.5.2 Organoid generation and sampling. Cells remaining after scRNA-seq allocation were initiated and maintained as patient-derived organoid cultures as previously described[25,40]. In brief, digested cells were seeded in 3-dimensional (3D) Growth-factor Reduced Matrigel (Corning) and fed with human complete organoid medium containing advanced DMEM/F12 (Gibco), 10 mM HEPES (Gibco), 1x GlutaMAX (Gibco), 500 nM A83-01 (Tocris), 50 ng/mL mEGF (Peprotech), 100 ng/mL mNoggin (Peprotech), 100 ng/mL hFGF10 (Peprotech), 10 nM hGastrin I (Sigma),

1.25 mM N-acetylcysteine (Sigma), 10 mM Nicotinamide (Sigma), 1x B27 supplement (Gibco), R-spondin1 conditioned media 10% final, Wnt3A conditioned media 50% final, 100 U/mL penicillin/streptomycin (Gibco), and 1x Primocin (Invivogen) (**Supplemental Table S3.6**). 10 μM Y27632 (Selleck) was included in the culture medium of newly initiated samples until the first media exchange. For propagation, organoids were dissociated with TrypLE (Gibco) before re-seeding into fresh Matrigel and culture medium.

After initial processing of fresh tissue specimens, we monitored samples closely for organoid growth. We did not passage organoids at set time intervals, as there was significant variability in the time needed to establish relatively robust growth of organoids (**Figure 3.4C**). Instead, we maintained early passage organoids until they reached relative confluence, and then passaged them at low split ratios (1:1, 1:1.5, or 1:2 dilutions) in complete organoid medium to promote continued growth. In one case, PANFR0489R, cells persisted as individuals and small organoids after initiation in complete organoid medium, but did not grow and expand cell numbers significantly. Approximately 15 weeks after initiation, we switched a portion of the surviving cells to organoid medium without A83-01 or mNoggin, and observed renewed growth of organoids under these media conditions but not of those that remained in complete organoid medium. Consequently, we expanded this sample in media without A83-01 or mNoggin, including performing early passage scRNA-seq. After several additional passages, once the organoids were robustly growing, we were able to transition back to complete organoid medium with no apparent change in growth rate, morphology, or transcriptional phenotype. All other serially sampled organoids were maintained and assessed by scRNA-seq in complete medium.

For scRNA-seq of organoid samples, we passaged organoids and allowed them to grow for 6 days before then dissociating, counting, and allocating 15,000 viable cells for Seq-Well. By standardizing the collection of organoid scRNA-seq samples at 6 days after passaging, we tried to minimize bias arising from cell cycle differences in samples at different degrees of confluence.

3.5.3 Testing organoid phenotypes under different matrix and media conditions. For adaptation of patient-derived organoids onto 2-dimensional (2D) culture surfaces as patient-derived cell lines, tissue culture plates were pre-coated with 100 μg/mL Matrigel dissolved in basal media for 2 hours at 37°C before washing with PBS. Established organoid models were dissociated and seeded onto these Matrigel-coated culture wells in complete organoid media. In parallel, a portion of these

passage-matched organoid cells were re-seeded into Matrigel droplets as above. Cells were cultured in both matrix conditions in complete organoid media until they were confluent, approximately 2-3 weeks. Cells were collected and lysed using Trizol before snap freezing. RNA was isolated and purified as described below ("Bulk RNA-sequencing of organoids" section) using chloroform extraction, aqueous phase isolation, and processing using the Qiagen AllPrep DNA/RNA/miRNA Universal kit before being submitted for sequencing.

For scRNA-seq assessment of organoid phenotypes when cultured under different media conditions, established organoid models were passaged as above by dissociating and reseeding into Matrigel droplets. A portion of the cells were cultured with complete organoid media ("Complete media"), while a distinct portion of passage-matched cells were cultured in "Minimal" media, which consisted of advanced DMEM/F12 (Gibco), 10 mM HEPES (Gibco), 1x GlutaMAX (Gibco), 100 U/mL penicillin/streptomycin (Gibco), and 1x Primocin (Invivogen) (**Supplemental Table S3.6**). Cells were cultured for 6 days before being collected, dissociated, and aliquoted for scRNA-seq. Images were taken with an Olympus XM10 camera mounted to an Olympus CKX41 microscope 1 day after seeding and again after 11 days in culture to assess organoid growth in both conditions. The portion of cells cultured in minimal media were maintained in the same conditions for a longer duration and harvested again for scRNA-seq at 27 days and 59 days after the initial introduction of minimal media. To mirror the standard scRNA-seq workflow, the cells harvested at the 27- and 59-day timepoints were collected 6 days after passaging.

In addition to the minimal media experiment, organoid cells were also cultured in standard cell line media ("RP10"), which contains RPMI-1640 (Gibco) and 100 U/mL penicillin/streptomycin (Gibco) with 10% fetal bovine serum (Sigma), or in reduced organoid media "OWRNA", which consists of advanced DMEM/F12 (Gibco), 10 mM HEPES (Gibco), 1x GlutaMAX (Gibco), 50 ng/mL mEGF (Peprotech), 100 ng/mL hFGF10 (Peprotech), 10 nM hGastrin I (Sigma), 1.25 mM N-acetylcysteine (Sigma), 10 mM Nicotinamide (Sigma), 1x B27 supplement (Gibco), 100 U/mL penicillin/streptomycin (Gibco), and 1x Primocin (Invivogen) (i.e. complete organoid medium with removal of WNT3A, RSPONDIN-1, NOGGIN, and A-8301; **Supplemental Table S3.6**). Furthermore, OWRNA reduced organoid medium served as the baseline control medium when assessing the effect of specific factors (IFNGγ and TGF-□1) from the TME on transcriptional phenotypes. Cells were cultured for 6 days before being collected, dissociated, and aliquoted for scRNA-seq in each of the following conditions: RP10, OWRNA,

OWRNA with 50 ng/mL IFNGγ (Peprotech), and OWRNA with 5 ng/mL TGFB1 (Peprotech) (**Supplemental Table S3.6**). The cells cultured under the IFNGγ and TGF-β1 conditions were maintained in culture and harvested again for scRNA-seq 38 days after being introduced to these new media conditions. For these longer duration timepoints, cells were again passaged 6 days before collecting for scRNA-seq.

3.5.4 Testing transcriptional phenotype changes in an established cell line under organoid media conditions.

For scRNA-seq assessment of transcriptional phenotypes of the established pancreatic cancer cell line CFPAC1 under different media conditions, CFPAC1 cells were cultured in parallel in either standard cell line medium RP10 or complete organoid medium. Cells were cultured for 6 days before being collected, dissociated, and aliquoted for scRNA-seq. Additionally, the CFPAC1 cells cultured under complete organoid medium were maintained in the same conditions and harvested again for scRNA-seq 33 days after the initial introduction of complete organoid medium. CFPAC1 cells grown in complete media for the later 33-day timepoint were collected 6 days after passaging, and media was refreshed 3 days after this final passage.

3.5.5 Single-cell RNA-seq (scRNA-seq) data library generation, sequencing, and alignment.

ScRNA-seq processing followed the Seq-Well protocol, uniquely compatible with low-input samples[38,39]. Briefly, arrays were preloaded with RNA capture beads (ChemGenes) and stored in quenching buffer until used. Prior to cell loading, arrays were resuspended in 5 mL RPMI-1640 medium with 10% fetal bovine serum (both from Gibco, hereafter referred to as RP10). After dissociation, single-cell suspensions were manually counted and diluted to 15,000 cells per 200 μL of RP10 when cell numbers allowed. Excess RP10 was aspirated from the array and cells were loaded onto arrays. Excess cells were washed off with PBS (4x5 mL, Gibco), briefly left in RPMI (5 mL) and cell+bead pairs were sealed for 40 minutes at 37°C using a polycarbonate membrane (Fisher Scientific NC1421644). Arrays were rocked in lysis buffer for 20 minutes and RNA was hybridized onto the beads for 40 minutes. Beads were removed and reverse transcription was performed overnight using Maxima H Minus Reverse Transcriptase (Thermo Fisher EP0753). Prior to sequencing, the beads underwent an exonuclease treatment (NewEngland Biolabs M0293L) and second strand synthesis *en masse* followed by whole transcriptome amplification

(WTA, Kapa Biosystems KK2602) in 1,500 bead reactions (50 μL). cDNA was isolated using Agencourt AMPure XP beads (Beckman Coulter, A63881) at 0.6X SPRI (solid-phase reversible immobilization) followed by a 1X SPRI and quantified using a Qubit dsDNA High Sensitivity assay kit (Thermo Fisher Q32854). Library preparation was performed using Nextera XT DNA tagmentation (Illumina FC-131-1096) and N700 and N500 indices specific to a given sample. Tagmented and amplified sequences were purified with a 0.6X SPRI. cDNA was loaded onto either an Illumina Nextseq (75 Cycle NextSeq500/550v2 kit) or Novaseq (100 Cycle NovaSeq6000S kit, Broad Institute Genomics Platform) at 2.4 pM. Regardless of platform, the paired end read structure was 21 bases (cell barcode and UMI) by 50 bases (transcriptomic information) with an 8 base pair (bp) custom read one primer. The demultiplex and alignment protocol was followed as previously described[64]. While Novaseq data were directly output as FASTQs, Nextseq BCL files were converted to FASTQs using bcl2fastq2. The resultant Nextseq and Novaseq FASTQs were demultiplexed by sample based on Nextera N700 and N500 indices. Reads were then aligned to the hg19 transcriptome using the cumulus/dropseq_tools pipeline on Terra maintained by the Broad Institute using standard settings.

3.5.6 Bulk RNA-sequencing of organoids. RNA was obtained for bulk RNA-sequencing from established organoids using one of two approaches. Dissociated organoids were resuspended into cold Matrigel, added as droplets to tissue culture plates (Greiner BioOne), and allowed to polymerize for 30 minutes before addition of media. Organoids were grown for 14-21 days (until confluent) under these conditions with regular media changes. At the time of harvest, cells were washed with cold phosphate buffered saline (PBS) at 4°C, then lysed with Trizol (Invitrogen) before snap-freezing. To isolate RNA, we performed chloroform extraction with isolation of the aqueous phase before processing RNA as per protocols outlined in the Qiagen AllPrep DNA/RNA/miRNA Universal kit.

In the second approach, dissociated organoids were resuspended in a solution of 10% Matrigel in complete organoid media (volume/volume) and cultured in ultra-low-attachment culture flasks (Corning). Organoids were grown for 14-21 days (until confluent) before pelleting, washing with cold PBS at 4°C until most Matrigel was dissipated, and then snap frozen. For RNA isolation, cell pellets were homogenized using buffer RLT Plus (Qiagen) and a Precellys homogenizer. Samples were then processed for both DNA extraction and RNA isolation as per the

Qiagen AllPrep DNA/RNA/miRNA Universal kit. Purified RNA was then submitted for sequencing by the Broad Institute Genomics Platform.

In brief, total RNA was quantified using the Quant-iT RiboGreen RNA Assay Kit (Thermo Fisher R11490) and normalized to 5 ng/μL. Following plating, 2 μL of a 1:1000 dilution of ERCC RNA controls (Thermo Fisher 4456740) were spiked into each sample. An aliquot of 200 ng for each sample was transferred into library preparation which uses an automated variant of the Illumina TruSeq Stranded mRNA Sample Preparation Kit. This method preserves strand orientation of the RNA transcript, and uses oligo dT beads to select mRNA from the total RNA sample followed by heat fragmentation and cDNA synthesis from the RNA template. The resultant 400 bp cDNA then goes through dual-indexed library preparation: 'A' base addition, adapter ligation using P7 adapters, and PCR enrichment using P5 adapters. After enrichment, the libraries were quantified using Quant-iT PicoGreen (1:200 dilution; Thermo Fisher P11496). After normalizing samples to 5 ng/μL, the set was pooled and quantified using the KAPA Library Quantification Kit for Illumina Sequencing Platforms. The entire process was performed in 96-well format and all pipetting was done by either Agilent Bravo or Hamilton Starlet.

Pooled libraries were normalized to 2 nM and denatured using 0.1 N NaOH prior to sequencing. Flowcell cluster amplification and sequencing were performed according to the manufacturer's protocols using the NovaSeq 6000. Each run was a 101 bp paired-end with an eight-base index barcode read. Data were analyzed using the Broad Picard Pipeline which includes de-multiplexing and data aggregation (https://broadinstitute.github.io/picard/). FASTQ files were then processed as described below (see Bulk RNA-sequencing analysis).

3.5.7 Multiplex immunofluorescence imaging. A multi-marker panel was developed to characterize tumor cell subtype in formalin-fixed paraffin-embedded (FFPE) 4μm tissue sections using multiplex immunofluorescence. The panel comprises markers associated with either a basal (Keratin-17: Thermo Fisher MA513539 and s100a2: Abcam 109494) or classical (cldn18.2: Abcam 241330, GATA6: CST 5851 and TFF1: Abcam 92377) subtype. Additionally, DAPI (Akoya Biosciences FP1490) was included for identification of nuclei and pan-cytokeratin (AE1/AE3: DAKO M3515; C11: CST 4545) for identification of epithelial cells. Secondary Opal Polymer HRP mouse and rabbit (ARH1001EA), Tyramide signal amplification and Opal fluorophores (Akoya Biosciences) were used to detect primary antibodies (Keratin-17, Opal 520;

s100a2, Opal 650; GATA6, Opal 540; cldn18.2, Opal 570; TFF1, Opal 690; panCK, Opal 620). Prior to use in multiplex staining, primary antibodies were first optimized via immunohistochemistry on control tissue to confirm contextual specificity. Monoplex immunofluorescence and iterative multiplex fluorescent staining were then used to optimize staining order, antibody-fluorophore assignments and fluorophore concentrations. Multiplex staining was performed using a Leica BOND RX Research Stainer (Leica Biosystems, Buffalo, IL) with sequential cycles of antigen retrieval, protein blocking, primary antibody incubation, secondary antibody incubation, and fluorescent labeling. Overview images of stained slides were acquired at 10X magnification using a Vectra 3.0 Automated Quantitative Imaging System (Perkin Elmer, Waltham, MA) and regions of interest (ROIs) were selected for multispectral image acquisition at 20X. After unmixing using a spectral library of single-color references, each image was inspected to ensure uniform staining quality and adequate tumor representation.

3.6 Data analysis

3.6.1 Mutation and CNV identification from bulk DNA-sequencing.

For targeted DNA-sequencing of clinical samples, next-generation sequencing using a custom-designed hybrid capture library preparation was performed on an Illumina HiSeq 2500 with 2x100 paired-end reads, as previously described (Garcia et al., 2017; Sholl et al., 2016). Sequence reads were aligned to reference sequence b37 edition from the Human Genome Reference Consortium using bwa, and further processed using Picard (version 1.90, http://broadinstitute.github.io/picard/) to remove duplicates and Genome Analysis Toolkit (GATK, version 1.6-5-g557da77) to perform localized realignment around indel sites. Single nucleotide variants were called using MuTect v1.1.45, insertions and deletions were called using GATK Indelocator. Copy number variants and structural variants were called using the internally-developed algorithms RobustCNV and BreaKmer followed by manual review[65]. RobustCNV calculates copy ratios by performing a robust linear regression against a panel of normal samples. The data were segmented using circular binary segmentation, and event identification was performed based on the observed variance of the data points (Bi et al., 2017).

We computed the cytoband-level copy number calls and weighted (by length) average segment means across the covered regions of each cytoband using ASCETS (Spurr et al., 2020). Briefly, cytobands were considered amplified/deleted if more than 70% of the covered regions had

a log2 copy ratio of greater than 0.2/less than -0.2, and were considered neutral if more than 70% of the covered regions had a log2 copy ratio between -0.2 and 0.2.

3.6.2 Single-cell data quality pre-processing and initial cell type discovery.

All single-cell data analysis was performed using the R language for Statistical Computing (v3.5.1). Each biopsy sample's digital gene expression (DGE) matrix (cells x genes) was trimmed to exclude low quality cells (<400 genes detected; <1,000 UMIs; >50% mitochondrial reads) before being merged together (preserving all unique genes) to create the larger biopsy dataset. The merged dataset was further trimmed to remove cells with >8,000 genes which represent outliers and likely doublet cells. We also removed genes that were not detected in at least 50 cells. The same metrics were applied to the organoid single-cell cohort (see below). On a per cell basis, UMI count data was divided by total transcripts captured and multiplied by a scaling factor of 10,000. These normalized values were then natural log transformed for downstream analysis (i.e. log-normalized cell x gene matrix). Initial exploration of the data was performed using the R package Seurat (v2.3.4) and followed two steps: 1) SNN-guided quality assessment and 2) cell type composition determination. In step 1, we intentionally left cells in the DGE matrix of dubious quality (e.g. % mitochondrial reads >25% but <50%), performed principal component analysis (PCA) over the variable genes (n = 1,070 genes), and input the first 50 PCs (determined by Jackstraw analysis implemented through Seurat) to build an SNN graph and cluster the cells (res = 1; k.param = 40). The inclusion of poor-quality cells essentially acts as a variance "sink" for other poor-quality cells and they cluster together based on their shared patterns in quality-associated gene expression. This method helped to nominate additional low quality (e.g. defined exclusively by mitochondrial genes) or likely doublet cells (e.g. clusters defined by co-expression of divergent lineage markers) which were removed from the dataset (n=1,678 cells). This led to an overall high-quality dataset of single-cells with a low overall faction of mitochondrial reads (median = 0.09) for downstream analysis (**Supplemental Figure S3.1B**)

Using the trimmed dataset, we proceeded to step 2 using a very similar workflow as above but with slightly altered input conditions for defining clusters. Here we used PCs 1-45 and their associated statistically significant genes for building the SNN graph and determining cluster membership (resolution = 1.2; k.param = 40). This identified the 36 clusters shown (visualized using *t*-SNE; perplexity, 40; iterations, 2,500) in **Supplemental Figure S3.1C**. The expression of

known markers was used to collapse clusters containing shared lineage information. For example, clusters 1, 2, and 4 all express high levels of macrophage markers—*CD14*, *FCGR3A (CD16)*, *CD68*—and were accordingly collapsed for this first pass analysis (**Supplemental Figure S3.1C,G**). To aid our cell type identification, we performed a ROC test implemented in Seurat to confirm the specificity (power > 0.6) of the top marker genes used to discern the cell types. Combined with inferred CNV information (see below), this analysis confirmed the presence of 11 broad non-malignant cell types in our biopsy dataset (**Supplemental Table S3.2**). Variation in the SNN graph parameters above did not strongly affect cell type identification.

3.6.3 Single-cell CNV identification.

To confirm the identity of the putative malignant clusters identified in **Supplemental Figure S3.1D**, we estimated single-cell CNVs as previously described by computing the average expression in a sliding window of 100 genes within each chromosome after sorting the detected genes by their chromosomal coordinates[3,13]. We used all T/NK, Fib, Hep, and Endo cells identified above as reference normal populations for this analysis. Complete information on the inferCNV workflow used for this analysis can be found here https://github.com/broadinstitute/inferCNV/wiki. To compare with bulk targeted DNA-sequencing, we collapsed individual probes to cytoband-level information (weighted average of log2 ratios across each cytoband, see above) within each sample. ScRNA-seq-inferred CNVs showed high concordance across samples with the bulk measurements and suggests that, at least by this metric, we are likely sampling the same dominant clones within sequential but distinct cores from each needle biopsy procedure (**Supplemental Figure S3.1E**). For plotting CNV profiles in putative malignant versus normal cells (**Supplemental Figure S3.1F**), we computed the average CNV signal for the top 5% of altered cells in each biopsy and correlated all cells in that biopsy to the averaged profile as has been previously described[7]. Relation of this correlation coefficient to the CNV score (mean square deviation from diploidy) in the single cells from each biopsy shows consistent separation of malignant from non-malignant cells, and, combined with membership in patient-specific SNN clusters, substantiates the identification of malignant cells in our dataset.

3.6.4 Subclonal analysis with single-cell inferred CNVs.

The inferCNV workflow can be used to call subclonal genetic variation with high sensitivity and is comprehensively outlined here https://github.com/broadinstitute/inferCNV/wiki [3,13,43]. Briefly, we used a six-state Hidden Markov Model (i6-HMM) to predict relative copy number status (complete loss to >3x gain) across putative altered regions in each cell. A Bayesian latent mixture model then evaluated the posterior probability that a given copy number alteration is a true positive. We set a relatively stringent cutoff for this step (BayesMaxPNormal = 0.2) to only include high probability alterations for downstream clustering. The results of this filtered i6-HMM output were then used to cluster the single cells using Ward's method. We used inferCNV's "random trees" method to test for statistical significance ($P < 0.05$, 100 random permutations for each split) at each tree bifurcation and only retained subclusters that had statistical evidence underlying the presumed heterogeneity. To track subclonal heterogeneity between biopsy and matched organoid cells in **Figure 3.3G** and **Supplemental Figure S3.5E-I**, the above workflow was implemented within each biopsy and the relevant matched organoid samples, essentially treating all cells as the same "tumor" and allowing the CNVs to determine cell sorting agnostic to sample-of-origin. The results of the HMM output can be used to infer gene-level information based on which genes are in the affected window. This (like the rest of the HMM workflow) is computed over groups of cells (e.g. samples or sub-clones) and used to map *KRAS* and other alterations to samples (**Figure 3.3A-F**) or sub-clones (**Figure 3.3G, Supplemental Figure S3.5E-I**).

3.6.5 Subclustering of malignant and non-malignant cells.

Detailed phenotyping required splitting the dataset into malignant and non-malignant fractions. After subsetting to only the malignant cells, we re-scaled the data and ran PCA including the first 35 PCs for SNN clustering and *t*-SNE visualization. This PCA was used to determine the PanNET identity for PANFR0580 (**Supplemental Figure S3.2A**). After removing PANFR0580, we repeated the steps above and used this new PCA for the remainder of PDAC malignant cell analysis. Subsequent phenotyping for malignant cells is discussed below (**Generation of expression signatures/scores**). A similar approach was used for calling the non-malignant subsets in **Figure 3.5A**. To determine the specific phenotypes within T/NK, macrophage, and mesenchymal populations, we separately subclustered these groups using PCs 1-20 and a resolution of 0.6 in each case. Of note, subclustering within the macrophages revealed a distinct cluster of cells co-expressing markers of both T/NK cells and macrophages (n=491 cells). We

discarded these cells as likely doublets, as have others, and re-ran the macrophage PCA and clustering[45,46]. These cells are included in the full dataset in case they are of interest to others. Each unbiased analysis helped to define the non-malignant phenotypes summarized in **Figures 3.5 & 3.6** and **Supplemental Figure S3.6**.

3.6.6 Generation of expression signatures/scores.

All expression scores were computed as previously described by taking a given input set of genes and comparing their average relative expression to that of a control set (n=100 genes) randomly sampled to mirror the expression distribution of the genes used for the input[13]. While all scores were computed in the same way, choosing the genes for input varied. We have outlined the relevant approaches below. Where correlations (Pearson's *r)* are performed over genes, we used the log-transformed UMI count data for each case. Unless otherwise noted, we selected the top 30 statistically significant genes for each signature (>3 s.d. above the mean for shuffled data) for visualization and scoring.

Cell cycle. We utilized previously established signatures for G1/S (n=43 genes) and G2/M (n=55 genes) to place each cell along this dynamic process (Tirosh et al., 2016a). After inspecting the distribution of scores in the complete dataset, we considered any cell >1.5 s.d. above the mean for either the G1/S or the G2/M scores to be cycling[8].

*Basal and classical programs***.** We started by scoring each malignant single cell for the basal-like and classical genes identified by Moffitt et al., 2015 as these were well described by unbiased analysis in our data (PCA, **Supplemental Figure S3.2B**). To determine programs associated with basal and classical phenotypes, we correlated the aforementioned basal and classical scores to the entire gene expression matrix containing malignant cells and selected the 1,909 genes significantly associated with either subtype (r > 0.1; >3 s.d. above the mean for shuffled data, full data in **Supplemental Table S3.3**). Biological pathway correlates for basal and classical mirrored previous work, and are summarized in **Supplemental Figure S3.3D**,**E**. For visualization, we use the "scCorr" basal and classical genes (top 30 correlated genes for each). We used these basal and classical scores to order the cells by their polarization or "score difference", simply the difference of the two scores, and revealed a significant fraction of cells co-expressing intermediate levels of both phenotypes (**Supplemental Figure S3.3A**,**B**).

*Intermediate transitional program***.** Intermediate cells showed associations with features across several additional PCs, but lacked a single dominant axis. To define a consensus set of genes that are preferentially expressed by cells in this intermediate state, we computed the Euclidean distance to the line representing equal basal and classical co-expression for each cell. To limit the influence of cell quality on this analysis and to specifically identify genes related to co-expression, we used cells from each group (basal, intermediate, and classical) with fractionally low mitochondrial genes (<0.2) and non-zero basal or classical expression (basal or classical score > 0) and correlated their Euclidean distance (**Supplemental Figure S3.3C**) to the entire gene expression matrix of malignant cells. Next, for each gene positively associated with this intermediate state (Pearson's r >0), we subtracted the second highest correlation coefficient for each subtype-associated gene (basal and classical), and then re-ranked the matrix by this corrected value. This enriched for genes more specific to the intermediate state by excluding those that were also associated with basal or classical programs. We then selected the 115 genes with a corrected correlation value >0.1 ($P< 0.00001$, shuffled data) as our intermediate transitional (IT) signature (**Supplemental Figure S3.3D**, **Supplemental Table S3.3**). Single cells were classified based on Euclidian distance where <0.2 are defined as intermediate transitional and the remainder (Euclidian distance >0.2) by their maximal of either basal or classical scores. We binned each organoid cell (e.g. **Figure 3.4B**,**C**) by its maximal expression for one of the 3 *in vivo* scores (basal, classical, or IT). Here a cell must be within 1 s.d. of the mean expression for a given subtype *in vivo*, else it was considered "organoid-specific" as this program was superimposed on all organoid cells, regardless of their subtype identity (**Figure 3.4B**). We used these classifications to summarize overall tumor composition and visualize the groups. Tumor heterogeneity measures were not significantly affected by changing these cutoffs.

Non-Malignant programs. TAM signatures were determined similar to above and previous work[3,45,46]. Using PCA as an anchor (**Supplemental Figure S3.6C**), we correlated expression within the TAM compartment to either *FCN1*, *SPP1*, or *C1QC* (top loaded genes on each relevant PC) and merged the resultant correlation coefficients for every detected gene to the 3 subtypes into one matrix (i.e. a 16,920 x 3 matrix). For each TAM type (i.e. vector of correlation coeffects to each marker), we first ranked the matrix by decreasing correlation coefficient, selected only the most significantly associated genes to that type (r > 0.1; >3 s.d. above the mean for shuffled data), subtracted the second highest correlation coefficient for each subtype-associated gene, and then

re-ranked the matrix by this corrected value. We repeated this procedure for each TAM subtype independently. This ensures that the genes selected are specific to a given TAM subset and do not describe general TAM features. The top 30 genes for each were used for scoring and visualization (**Supplemental Table S3.2**; **Supplemental Figure S3.6D**).

CAF phenotypes were determined using a similar workflow. To examine fibroblast heterogeneity, we removed a subset of adrenal endocrine cells (cluster 4, 40 cells; **Figure 3.5C**) and then performed PCA of mesenchymal cells. PC1 was driven by spillover genes (likely contributed from ambient RNA) and lacked any coherent biological program and was not considered further. PCs 2 and 3 by contrast where consistent with variable mesenchymal (PC2) and inflammatory (PC3) CAF phenotypes. All these cells scored highly for previous myCAF gene expression programs so this phenotype did not fully explain the heterogeneity in mesenchymal cells, but did suggest their identity as CAFs. Again, using correlation, we determined the genes driving low PC2 scores (Dermal-like), and high PC2 scores (Pericyte-like), as well as those associated with the high PC3 scores (Inflammatory). As before, we used the top 30 genes for each subset scoring and visualization. These same genes (Dermal-like and Pericyte-like) were used to examine bulk RNA-seq profiles and their difference in each sample quantifies which phenotype is favored in the bulk averages (**Figure 3.5F**).

3.6.7 TME associations.

We determined the transcriptional-subtype-dependent composition of the TME (**Figure 3.6A-C**) following two steps. First, we computed the Simpson's Index (measure of ecological diversity) using the count of each non-malignant cell type captured from each sample as input (**Figure 3.6A**,**B**) and correlated each biopsy's diversity score to its basal vs. classical commitment score. Importantly, the number of non-malignant cells captured from each biopsy was not associated with basal vs. classical commitment score ($r = 0.09$). Next, to understand which cell types drive these differences, we computed the fractional representation for every non-malignant cell type in each core needle biopsy and determined their pairwise correlation distance (Pearson's *r*) followed by hierarchical clustering using Ward's method (dendrogram in **Figure 3.6B**). For both of these analyses we only used samples with >200 non-malignant cells captured (**Supplemental Figure S3.6L**).

3.6.8 Matched organoid clustering and cell-typing.

After applying similar quality metrics as above, we performed PCA, SNN clustering, and *t*-SNE embedding for 31,867 cells including organoid cells and all malignant cells from primary PDAC biopsies (PCs 1-50; resolution=1.2; k.param=45; perplexity=45; max_iter=2,500), and identified 39 total clusters. Organoids clustered separately from their matched biopsies, suggesting expression and/or CNV related drift in culture. Only two SNN clusters—clusters 4 and 32—were admixed by sample. We determined the specific gene expression programs in these two clusters via differential expression testing by Wilcoxon rank sum test ($P < 0.05$, Bonferroni correction; log(fold change) > 0.5). These comparisons were done in a "1 versus rest" fashion, testing for genes defining each cluster (4 or 32) compared to the entire dataset. Their expression profiles were consistent with fibroblasts (cluster 32) and epithelial cells (cluster 4; **Supplemental Figure S3.5B**,**C**).

3.6.9 Correlation distances for genotype and phenotype.

To generate correlation distances for genotype and phenotype, each single cell in a biopsy-organoid pair was represented by two vectors of information: (i) a phenotype vector containing expression values for basal and classical genes (scCorr basal and classical genes, n = 60 genes) and (ii) a genotype vector containing the average CNV score for each cytoband. The phenotype and genotype distances between every single cell within a biopsy/early organoid pair was computed from these vectors using a correlation-based (Pearson's *r*) distance metric of the form $d = (1-r)/2$. This resulted in two distance matrices of n x n dimension where n is the total number of cells from each biopsy/early organoid sample pair. Values in **Figure 3.4A** are computed by averaging the values for *d* between only early organoid and matched biopsy cells.

3.6.10 Matched biopsy vs. organoid malignant cell comparison.

For CNV-confirmed malignant cells from each biopsy and its matched organoid (earliest passage), we used differential expression (Wilcoxon rank sum test; $P < 0.05$, Bonferroni correction; log(fold change) > 0.3) to understand the features lost from malignant cells in the *in vivo* setting and gained when transitioning into growth in organoid culture. We required any gene to be significantly differentially expressed in at least 3 model-biopsy comparisons to summarize the consistent

changes. We repeated this same workflow for both organoid- and biopsy-specific genes (**Supplemental Table S3.5**) outlined in **Figure 3.4B** and **Figure 3.4D-F**, respectively.

3.6.11 Biopsy paracrine and autocrine subtype-specific factor analysis.

Factors present in the TME but absent from organoid culture could originate from at least two sources, the tumor cells themselves (autocrine) or non-tumor cells in the local microenvironment (paracrine). We examined any gene with gene ontology annotations related to "cytokines", "chemokines", or "growth factors" and took the union of these lists, yielding 321 genes, 218 of which were detected in our dataset. For "autocrine" factors we performed differential expression between malignant cells binned as basal and classical, and then IT vs rest. A gene was considered differentially expressed if it passed a $P < 0.05$ with Bonferroni correction and a log(fold change) > 0.2 in one of these comparisons. Genes were then assigned to subtypes based on the log fold change direction (**Figure 3.7D**, **Supplemental Table S3.7**). Paracrine factors were determined in a similar manner with slight modifications. We grouped non-tumor cells into basal, classical or IT based on the average expression and clustering for malignant programs from their respective tumor samples (**Supplemental Figure S3.3G,H**). We then assessed for differential expression between all cells from a given group and the rest using the same cutoffs as above and sorted factors into subtypes based on their log fold change directionality (**Figure 3.7F**, **Supplemental Table S3.7**). We then visualized which cell type contributed the highest average expression for each factor in the cell types from the respective TMEs (**Figure 3.7G**).

3.6.12 Bulk RNA-sequencing analysis.

FASTQs for bulk RNA expression profiles were downloaded from the relevant repository (TCGA, https://toil.xenahubs.net; PDAC Cell lines, https://portals.broadinstitute.org/ccle), available in-house (Panc-Seq, metastatic PDAC), or generated for this study (organoid cohort)[15,18,53,55]. All were processed using the same pipeline. Briefly, each sample's sequences were marked for duplicates and then mapped to hg38 using STAR. After running QC checks using RNAseqQC, gene-level count matrices were generated using RSEM. Instructions to run the pipeline are given in the Broad CCLE github repository https://github.com/broadinstitute/ccle_processing. Length-normalized values (TPM) were then transformed according to log2(TPM+1) for downstream analysis. The entire dataset was scaled and centered to allow relative comparisons across sample

types (e.g. tumors, organoids, and cell lines). Signature scores were computed as above (e.g. basal and classical; see **Generation of expression signatures/scores** above)[4].

3.6.13 Tumor phenotyping from mIF data.

Supervised machine learning algorithms were applied for tissue and cell segmentation (inForm 2.4.1, Akoya Biosciences). Single-cell-level imaging data were exported and further processed and analyzed using R (v3.6.2). To assign phenotypes to individual tumor epithelial cells, mean expression intensity in the relevant subcellular compartment was first used to classify cells as positive or negative for each of the 5 markers. Combinatorial expression patterns for the five markers were then used to phenotypically classify cells as basal, classical, co-expressing / IT or marker negative (3 combinations of 2 basal markers, 7 combinations of 3 classical markers, 1 pan-marker negative, 21 combinations of co-expression of basal and classical markers, **Supplemental Figure S3.4A, Supplemental Table S3.4**). Tumor subtype composition was assessed by calculating the fraction of total tumor cells positive for each cell phenotype (**Supplemental Figure S3.4B**, excluding pan-marker negative cells).

3.7 Reference

1. Hyman, D.M., Taylor, B.S., and Baselga, J. (2017). Implementing Genome-Driven Oncology. Cell *168*, 584-599.
2. Kim, C., Gao, R., Sei, E., Brandt, R., Hartman, J., Hatschek, T., Crosetto, N., Foukakis, T., and Navin, N.E. (2018). Chemoresistance Evolution in Triple-Negative Breast Cancer Delineated by Single-Cell Sequencing. Cell *173*, 879-893 e813.
3. Patel, A.P., Tirosh, I., Trombetta, J.J., Shalek, A.K., Gillespie, S.M., Wakimoto, H., Cahill, D.P., Nahed, B.V., Curry, W.T., Martuza, R.L.*, et al.* (2014). Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma. Science *344*, 1396-1401.
4. Puram, S.V., Tirosh, I., Parikh, A.S., Patel, A.P., Yizhak, K., Gillespie, S., Rodman, C., Luo, C.L., Mroz, E.A., Emerick, K.S.*, et al.* (2017). Single-Cell Transcriptomic Analysis of Primary and Metastatic Tumor Ecosystems in Head and Neck Cancer. Cell *171*, 1611-1624 e1624.
5. Sade-Feldman, M., Yizhak, K., Bjorgaard, S.L., Ray, J.P., de Boer, C.G., Jenkins, R.W., Lieb, D.J., Chen, J.H., Frederick, D.T., Barzily-Rokni, M.*, et al.* (2019). Defining T Cell States Associated with Response to Checkpoint Immunotherapy in Melanoma. Cell *176*, 404.

6. Suva, M.L., and Tirosh, I. (2019). Single-Cell RNA Sequencing in Cancer: Lessons Learned and Emerging Challenges. Mol Cell *75*, 7-12.

7. Tirosh, I., Izar, B., Prakadan, S.M., Wadsworth, M.H., 2nd, Treacy, D., Trombetta, J.J., Rotem, A., Rodman, C., Lian, C., Murphy, G.*, et al.* (2016). Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-seq. Science *352*, 189-196.

8. van Galen, P., Hovestadt, V., Wadsworth Ii, M.H., Hughes, T.K., Griffin, G.K., Battaglia, S., Verga, J.A., Stephansky, J., Pastika, T.J., Lombardi Story, J.*, et al.* (2019). Single-Cell RNA-Seq Reveals AML Hierarchies Relevant to Disease Progression and Immunity. Cell *176*, 1265-1281 e1224.

9. Venteicher, A.S., Tirosh, I., Hebert, C., Yizhak, K., Neftel, C., Filbin, M.G., Hovestadt, V., Escalante, L.E., Shaw, M.L., Rodman, C.*, et al.* (2017). Decoupling genetics, lineages, and microenvironment in IDH-mutant gliomas by single-cell RNA-seq. Science *355*.

10. Filbin, M.G., Tirosh, I., Hovestadt, V., Shaw, M.L., Escalante, L.E., Mathewson, N.D., Neftel, C., Frank, N., Pelton, K., Hebert, C.M.*, et al.* (2018). Developmental and oncogenic programs in H3K27M gliomas dissected by single-cell RNA-seq. Science *360*, 331-335.

11. Hovestadt, V., Smith, K.S., Bihannic, L., Filbin, M.G., Shaw, M.L., Baumgartner, A., DeWitt, J.C., Groves, A., Mayr, L., Weisman, H.R.*, et al.* (2019). Resolving medulloblastoma cellular architecture by single-cell genomics. Nature *572*, 74-79.

12. Neftel, C., Laffy, J., Filbin, M.G., Hara, T., Shore, M.E., Rahme, G.J., Richman, A.R., Silverbush, D., Shaw, M.L., Hebert, C.M.*, et al.* (2019). An Integrative Model of Cellular States, Plasticity, and Genetics for Glioblastoma. Cell *178*, 835-849 e821.

13. Tirosh, I., Venteicher, A.S., Hebert, C., Escalante, L.E., Patel, A.P., Yizhak, K., Fisher, J.M., Rodman, C., Mount, C., Filbin, M.G.*, et al.* (2016b). Single-cell RNA-seq supports a developmental hierarchy in human oligodendroglioma. Nature *539*, 309-313.

14. Nam, A.S., Chaligne, R., and Landau, D.A. (2021). Integrating genetic and non-genetic determinants of cancer evolution by single-cell multi-omics. Nat Rev Genet *22*, 3-18.

15. Aguirre, A.J., Nowak, J.A., Camarda, N.D., Moffitt, R.A., Ghazani, A.A., Hazar-Rethinam, M., Raghavan, S., Kim, J., Brais, L.K., Ragon, D.*, et al.* (2018). Real-time Genomic Characterization of Advanced Pancreatic Cancer to Enable Precision Medicine. Cancer Discov *8*, 1096-1111.

16. Aung, K.L., Fischer, S.E., Denroche, R.E., Jang, G.H., Dodd, A., Creighton, S., Southwood, B., Liang, S.B., Chadwick, D., Zhang, A.*, et al.* (2018). Genomics-Driven Precision Medicine for Advanced Pancreatic Cancer: Early Results from the COMPASS Trial. Clin Cancer Res *24*, 1344-1354.

17. Bailey, P., Chang, D.K., Nones, K., Johns, A.L., Patch, A.M., Gingras, M.C., Miller, D.K., Christ, A.N., Bruxner, T.J., Quinn, M.C.*, et al.* (2016). Genomic analyses identify molecular subtypes of pancreatic cancer. Nature *531*, 47-52.

18. Cancer Genome Atlas Research Network (2017). Integrated Genomic Characterization of Pancreatic Ductal Adenocarcinoma. Cancer Cell *32*, 185-203 e113.

19. Chan-Seng-Yue, M., Kim, J.C., Wilson, G.W., Ng, K., Figueroa, E.F., O'Kane, G.M., Connor, A.A., Denroche, R.E., Grant, R.C., McLeod, J.*, et al.* (2020). Transcription phenotypes of pancreatic cancer are driven by genomic events during tumor evolution. Nat Genet *52*, 231-240.

20. Collisson, E.A., Sadanandam, A., Olson, P., Gibb, W.J., Truitt, M., Gu, S., Cooc, J., Weinkle, J., Kim, G.E., Jakkula, L.*, et al.* (2011). Subtypes of pancreatic ductal adenocarcinoma and their differing responses to therapy. Nat Med *17*, 500-503.

21. Connor, A.A., Denroche, R.E., Jang, G.H., Lemire, M., Zhang, A., Chan-Seng-Yue, M., Wilson, G., Grant, R.C., Merico, D., Lungu, I.*, et al.* (2019). Integration of Genomic and Transcriptional Features in Pancreatic Cancer Reveals Increased Cell Cycle Progression in Metastases. Cancer Cell *35*, 267-282 e267.

22. Moffitt, R.A., Marayati, R., Flate, E.L., Volmar, K.E., Loeza, S.G., Hoadley, K.A., Rashid, N.U., Williams, L.A., Eaton, S.C., Chung, A.H.*, et al.* (2015). Virtual microdissection identifies distinct tumor- and stroma-specific subtypes of pancreatic ductal adenocarcinoma. Nat Genet *47*, 1168-1178.

23. O'Kane, G.M., Grunwald, B.T., Jang, G.H., Masoomian, M., Picardo, S., Grant, R.C., Denroche, R.E., Zhang, A., Wang, Y., Lam, B.*, et al.* (2020). GATA6 Expression Distinguishes Classical and Basal-like Subtypes in Advanced Pancreatic Cancer. Clin Cancer Res.

24. Porter, R.L., Magnus, N.K.C., Thapar, V., Morris, R., Szabolcs, A., Neyaz, A., Kulkarni, A.S., Tai, E., Chougule, A., Hillis, A.*, et al.* (2019). Epithelial to mesenchymal plasticity and differential response to therapies in pancreatic ductal adenocarcinoma. Proc Natl Acad Sci U S A.

25. Tiriac, H., Belleau, P., Engle, D.D., Plenker, D., Deschenes, A., Somerville, T.D.D., Froeling, F.E.M., Burkhart, R.A., Denroche, R.E., Jang, G.H.*, et al.* (2018). Organoid Profiling Identifies Common Responders to Chemotherapy in Pancreatic Cancer. Cancer Discov *8*, 1112-1129.

26. Hayashi, A., Fan, J., Chen, R., Ho, Y.-j., Makohon-Moore, A.P., Lecomte, N., Zhong, Y., Hong, J., Huang, J., Sakamoto, H.*, et al.* (2020). A unifying paradigm for transcriptional heterogeneity and squamous features in pancreatic ductal adenocarcinoma. Nature Cancer *1*, 59-74.

27. Schleger, C., Verbeke, C., Hildenbrand, R., Zentgraf, H., and Bleyl, U. (2002). c-MYC activation in primary and metastatic ductal adenocarcinoma of the pancreas: incidence, mechanisms, and clinical significance. Mod Pathol *15*, 462-469.

28. Miyabayashi, K., Baker, L.A., Deschênes, A., Traub, B., Caligiuri, G., Plenker, D., Alagesan, B., Belleau, P., Li, S., Kendall, J.*, et al.* (2020). Intraductal Transplantation Models of Human Pancreatic Ductal Adenocarcinoma Reveal Progressive Transition of Molecular Subtypes. Cancer Discovery *10*, 1566-1589.

29. Muzumdar, M.D., Chen, P.Y., Dorans, K.J., Chung, K.M., Bhutkar, A., Hong, E., Noll, E.M., Sprick, M.R., Trumpp, A., and Jacks, T. (2017). Survival of pancreatic cancer cells lacking KRAS function. Nat Commun *8*, 1090.

30. Siegel, R.L., Miller, K.D., and Jemal, A. (2020). Cancer statistics, 2020. CA Cancer J Clin *70*, 7-30.

31. Balachandran, V.P., Beatty, G.L., and Dougan, S.K. (2019). Broadening the Impact of Immunotherapy to Pancreatic Cancer: Challenges and Opportunities. Gastroenterology *156*, 2056-2072.

32. Bernard, V., Semaan, A., Huang, J., San Lucas, F.A., Mulu, F.C., Stephens, B.M., Guerrero, P.A., Huang, Y., Zhao, J., Kamyabi, N.*, et al.* (2019). Single-Cell Transcriptomics of Pancreatic Cancer Precursors Demonstrates Epithelial and Microenvironmental Heterogeneity as an Early Event in Neoplastic Progression. Clin Cancer Res *25*, 2194-2205.

33. Biffi, G., Oni, T.E., Spielman, B., Hao, Y., Elyada, E., Park, Y., Preall, J., and Tuveson, D.A. (2019). IL1-Induced JAK/STAT Signaling Is Antagonized by TGFbeta to Shape CAF Heterogeneity in Pancreatic Ductal Adenocarcinoma. Cancer Discov *9*, 282-301.

34. Elyada, E., Bolisetty, M., Laise, P., Flynn, W.F., Courtois, E.T., Burkhart, R.A., Teinor, J.A., Belleau, P., Biffi, G., Lucito, M.S.*, et al.* (2019). Cross-Species Single-Cell Analysis of Pancreatic

Ductal Adenocarcinoma Reveals Antigen-Presenting Cancer-Associated Fibroblasts. Cancer Discov *9*, 1102-1123.

35. Ligorio, M., Sil, S., Malagon-Lopez, J., Nieman, L.T., Misale, S., Di Pilato, M., Ebright, R.Y., Karabacak, M.N., Kulkarni, A.S., Liu, A.*, et al.* (2019). Stromal Microenvironment Shapes the Intratumoral Architecture of Pancreatic Cancer. Cell *178*, 160-175 e127.

36. Ohlund, D., Handly-Santana, A., Biffi, G., Elyada, E., Almeida, A.S., Ponz-Sarvise, M., Corbo, V., Oni, T.E., Hearn, S.A., Lee, E.J.*, et al.* (2017). Distinct populations of inflammatory fibroblasts and myofibroblasts in pancreatic cancer. J Exp Med *214*, 579-596.

37. Ho, W.J., Jaffee, E.M., and Zheng, L. (2020). The tumour microenvironment in pancreatic cancer - clinical challenges and opportunities. Nat Rev Clin Oncol *17*, 527-540.

38. Gierahn, T.M., Wadsworth, M.H., 2nd, Hughes, T.K., Bryson, B.D., Butler, A., Satija, R., Fortune, S., Love, J.C., and Shalek, A.K. (2017). Seq-Well: portable, low-cost RNA sequencing of single cells at high throughput. Nat Methods *14*, 395-398.

39. Hughes, T.K., Wadsworth, M.H., 2nd, Gierahn, T.M., Do, T., Weiss, D., Andrade, P.R., Ma, F., de Andrade Silva, B.J., Shao, S., Tsoi, L.C.*, et al.* (2020). Second-Strand Synthesis-Based Massively Parallel scRNA-Seq Reveals Cellular States and Molecular Features of Human Inflammatory Skin Pathologies. Immunity *53*, 878-894 e877.

40. Boj, S.F., Hwang, C.I., Baker, L.A., Chio, II, Engle, D.D., Corbo, V., Jager, M., Ponz-Sarvise, M., Tiriac, H., Spector, M.S.*, et al.* (2015). Organoid models of human and mouse ductal pancreatic cancer. Cell *160*, 324-338.

41. Groger, C.J., Grubinger, M., Waldhor, T., Vierlinger, K., and Mikulits, W. (2012). Meta-analysis of gene expression signatures defining the epithelial to mesenchymal transition during cancer progression. PLoS One *7*, e51136..

42. Qadir, M.M.F., Alvarez-Cubela, S., Klein, D., van Dijk, J., Muniz-Anquela, R., Moreno-Hernandez, Y.B., Lanzoni, G., Sadiq, S., Navarro-Rubio, B., Garcia, M.T.*, et al.* (2020). Single-cell resolution analysis of the human pancreatic ductal progenitor cell niche. Proc Natl Acad Sci U S A *117*, 10876-10887.

43. Fan, J., Lee, H.O., Lee, S., Ryu, D.E., Lee, S., Xue, C., Kim, S.J., Kim, K., Barkas, N., Park, P.J.*, et al.* (2018). Linking transcriptional and genetic tumor heterogeneity through allele analysis of single-cell RNA-seq data. Genome Res *28*, 1217-1227.

44. Somerville, T.D.D., Xu, Y., Miyabayashi, K., Tiriac, H., Cleary, C.R., Maia-Silva, D., Milazzo, J.P., Tuveson, D.A., and Vakoc, C.R. (2018). TP63-Mediated Enhancer Reprogramming Drives the Squamous Subtype of Pancreatic Ductal Adenocarcinoma. Cell Rep *25*, 1741-1755 e1747.

45. Zhang, L., Li, Z., Skrzypczynska, K.M., Fang, Q., Zhang, W., O'Brien, S.A., He, Y., Wang, L., Zhang, Q., Kim, A.*, et al.* (2020). Single-Cell Analyses Inform Mechanisms of Myeloid-Targeted Therapies in Colon Cancer. Cell *181*, 442-459 e429.

46. Zilionis, R., Engblom, C., Pfirschke, C., Savova, V., Zemmour, D., Saatcioglu, H.D., Krishnan, I., Maroni, G., Meyerovitz, C.V., Kerwin, C.M.*, et al.* (2019). Single-Cell Transcriptomics of Human and Mouse Lung Cancers Reveals Conserved Myeloid Populations across Individuals and Species. Immunity *50*, 1317-1334 e1310.

47. Ascension, A.M., Fuertes-Alvarez, S., Ibanez-Sole, O., Izeta, A., and Arauzo-Bravo, M.J. (2020). Human Dermal Fibroblast Subpopulations Are Conserved across Single-Cell RNA Sequencing Studies. J Invest Dermatol.

48. Bartoschek, M., Oskolkov, N., Bocci, M., Lovrot, J., Larsson, C., Sommarin, M., Madsen, C.D., Lindgren, D., Pekar, G., Karlsson, G.*, et al.* (2018). Spatially and functionally distinct subclasses of breast cancer-associated fibroblasts revealed by single cell RNA sequencing. Nat Commun *9*, 5150.

49. Di Carlo, S.E., and Peduto, L. (2018). The perivascular origin of pathological fibroblasts. J Clin Invest *128*, 54-63.

50. Hosaka, K., Yang, Y., Seki, T., Fischer, C., Dubey, O., Fredlund, E., Hartman, J., Religa, P., Morikawa, H., Ishii, Y.*, et al.* (2016). Pericyte-fibroblast transition promotes tumor growth and metastasis. Proc Natl Acad Sci U S A *113*, E5618-5627.

51. Pelon, F., Bourachot, B., Kieffer, Y., Magagna, I., Mermet-Meillon, F., Bonnet, I., Costa, A., Givel, A.M., Attieh, Y., Barbazan, J.*, et al.* (2020). Cancer-associated fibroblast heterogeneity in axillary lymph nodes drives metastases in breast cancer through complementary mechanisms. Nat Commun *11*, 404.

52. Philippeos, C., Telerman, S.B., Oules, B., Pisco, A.O., Shaw, T.J., Elgueta, R., Lombardi, G., Driskell, R.R., Soldin, M., Lynch, M.D.*, et al.* (2018). Spatial and Single-Cell Transcriptional Profiling Identifies Functionally Distinct Human Dermal Fibroblast Subpopulations. J Invest Dermatol *138*, 811-825.

53. Cancer Genome Atlas Research, N., Weinstein, J.N., Collisson, E.A., Mills, G.B., Shaw, K.R., Ozenberger, B.A., Ellrott, K., Shmulevich, I., Sander, C., and Stuart, J.M. (2013). The Cancer Genome Atlas Pan-Cancer analysis project. Nat Genet *45*, 1113-1120.

54. Barretina, J., Caponigro, G., Stransky, N., Venkatesan, K., Margolin, A.A., Kim, S., Wilson, C.J., Lehar, J., Kryukov, G.V., Sonkin, D.*, et al.* (2012). The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. Nature *483*, 603-607.

55. Ghandi, M., Huang, F.W., Jane-Valbuena, J., Kryukov, G.V., Lo, C.C., McDonald, E.R., 3rd, Barretina, J., Gelfand, E.T., Bielski, C.M., Li, H.*, et al.* (2019). Next-generation characterization of the Cancer Cell Line Encyclopedia. Nature *569*, 503-508.

56. Alonso-Curbelo, D., Ho, Y.J., Burdziak, C., Maag, J.L.V., Morris, J.P.t., Chandwani, R., Chen, H.A., Tsanov, K.M., Barriga, F.M., Luan, W.*, et al.* (2021). A gene-environment-induced epigenetic program initiates tumorigenesis. Nature *590*, 642-648.

57. Li, J., Byrne, K.T., Yan, F., Yamazoe, T., Chen, Z., Baslan, T., Richman, L.P., Lin, J.H., Sun, Y.H., Rech, A.J.*, et al.* (2018). Tumor Cell-Intrinsic Factors Underlie Heterogeneity of Immune Cell Infiltration and Response to Immunotherapy. Immunity *49*, 178-193 e177.

58. Ben-David, U., Ha, G., Tseng, Y.Y., Greenwald, N.F., Oh, C., Shih, J., McFarland, J.M., Wong, B., Boehm, J.S., Beroukhim, R.*, et al.* (2017). Patient-derived xenografts undergo mouse-specific tumor evolution. Nat Genet *49*, 1567-1575.

59. Ben-David, U., Siranosian, B., Ha, G., Tang, H., Oren, Y., Hinohara, K., Strathdee, C.A., Dempster, J., Lyons, N.J., Burns, R.*, et al.* (2018). Genetic and transcriptional evolution alters cancer cell line drug response. Nature *560*, 325-330.

60. Benci, J.L., Xu, B., Qiu, Y., Wu, T.J., Dada, H., Twyman-Saint Victor, C., Cucolo, L., Lee, D.S.M., Pauken, K.E., Huang, A.C.*, et al.* (2016). Tumor Interferon Signaling Regulates a Multigenic Resistance Program to Immune Checkpoint Blockade. Cell *167*, 1540-1554 e1512.

61. Sahai, E., Astsaturov, I., Cukierman, E., DeNardo, D.G., Egeblad, M., Evans, R.M., Fearon, D., Greten, F.R., Hingorani, S.R., Hunter, T.*, et al.* (2020). A framework for advancing our understanding of cancer-associated fibroblasts. Nat Rev Cancer *20*, 174-186.

62. Fabre, M., Ferrer, C., Dominguez-Hormaetxe, S., Bockorny, B., Murias, L., Seifert, O., Eisler, S.A., Kontermann, R.E., Pfizenmaier, K., Lee, S.Y.*, et al.* (2020). OMTX705, a Novel FAP-Targeting ADC Demonstrates Activity in Chemotherapy and Pembrolizumab-Resistant Solid Tumor Models. Clin Cancer Res *26*, 3420-3430.

63. Ben-David, U., Beroukhim, R., and Golub, T.R. (2019). Genomic evolution of cancer models: perils and opportunities. Nat Rev Cancer *19*, 97-109.

64. Macosko, E. Z., Basu, A., Satija, R., Nemesh, J., Shekhar, K., Goldman, M., Tirosh, I., Bialas, A. R., Kamitaki, N., Martersteck, E. M., Trombetta, J. J., Weitz, D. A., Sanes, J. R., Shalek, A. K., Regev, A., & McCarroll, S. A. (2015). Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets. *Cell*, *161*(5), 1202–1214.

65. Abo, R. P., Ducar, M., Garcia, E. P., Thorner, A. R., Rojas-Rudilla, V., Lin, L., Sholl, L. M., Hahn, W. C., Meyerson, M., Lindeman, N. I., Van Hummelen, P., & MacConaill, L. E. (2015). BreaKmer: detection of structural variation in targeted massively parallel sequencing data using kmers. *Nucleic acids research*, *43*(3), e19.

Chapter 4: Conclusions

4.1 Motivations

I joined the Shalek Lab with the goal of collaborating with clinicians on projects with medically meaningful outcomes. Chapters 2 and 3 in this thesis outline the two projects that fulfilled that goal during my graduate studies. The unifying technology, scRNA-seq, has powered my work to draw high-resolution, biological insights across diverse disease contexts[1,2,3,4]. Based on our COVID-19 and PDAC results, we hope to build upon and disseminate these novel clinical research pipelines beyond the scope of our lab[5]. To this end, I discuss several alternative research areas where the utility and breadth of our findings can be demonstrated. Below, I summarize the broad conclusions of my work, future biological directions, and where technology is limiting, but ripe for innovation.

4.2 Building on the PDAC study

Our PDAC research used a novel clinical pipeline to define some of the rules governing tumor cell plasticity, and how malignant cells respond to specific perturbations in the microenvironment. Based on these findings, we tested drug efficacy in an isogenic, state-specific manner. In the future we intend to apply the principles outlined here to other tumor environments. Indeed, we are already processing banked tissues samples from diverse tumor types (some with basal-classical-like axes, others without) to expand on our hypotheses about RNA-state-dependent drug response more broadly. Early results suggest similar state-specific microenvironments that will hopefully show distinct therapeutic vulnerabilities akin to our initial PDAC findings[3]. Ultimately, the purpose of these high-fidelity models is to empower accurate drug discovery pipelines in PDAC and other intractable diseases. To that end, we continue to refine our organoid growth protocols to model tumor states beyond PDAC.

Our most mature effort inspired by our PDAC findings looks at tumor cell plasticity at minimal residual disease (MRD) in acute lymphoblastic leukemia (ALL). MRD poses a vexing barrier to cures across a range of human cancers, and its sparsity *in situ* makes its characterization challenging[6,7,8]. We sought to overcome the limitations of this low-input heterogeneous tissue to define cellular adaptations to oncogene withdrawal *in vivo* using a platform, analogous to that applied in PDAC, that links the single-cell functional and molecular profiling of primary human

tumors[1,2,9,10]. We tested this approach in the context of BCR-ABL-rearranged ALL, an archetype of oncogene-addicted tumors against which targeted kinase inhibitors with progressively increased potency and breadth of activity have improved complete response rates, but ultimately fail to eradicate MRD and cure patients[11]. Using our pipeline, we observed recurrent transcriptional adaptations in human leukemia cells that had survived sustained oncogene withdrawal, including stress response via p38 MAPK, upregulation of pre-B cell receptor (pre-BCR) signaling, and TNFα/NF-κB-mediated quiescence. These cellular programs enabled the eventual expansion of subclones harboring high-level single and compound resistance mutations that reactivated divergent oncogenic signaling through either STAT5 or ERK. While mutations in both pathways were detected at MRD, outgrowth of these clones in progression was associated with B-cell differentiation states, whose precise classification was enhanced through whole transcriptome-derived single-cell analysis. Incorporation of inhibitors of individual tumor's cellular adaptations within MRD improved the depth of *in vivo* responses compared to regimens targeting common resistance mutations alone. These data justify further preclinical development of MRD-targeted therapy and warrant further application of these pre-clinical pipelines more broadly.

4.3 Building models in SARS-CoV-2

Our research into the SARS-CoV-2 infection identified gene expression patterns in severe disease but was not powered to comprehensively test nominated avenues for therapeutic intervention. Given those data, and our experience in developing and benchmarking model systems, we aim to establish *in vivo* and *ex vivo* model systems that represent a spectrum of disease severity and allow for tractable, productive drug screening.

Going forward we have expanded our research to include non-human primates (NHPs), hamsters and organ-on-chip models[5]. Our NHP models generate a human-like immune response, though we see a less severe disease burden than that found in our patient cohort in Chapter 2. As such, in tandem, we are working with a hamster system that tends to be more representative of severe disease to better match our existing dataset. Finally, organ-on-chip models allow for tractable drug screening experiments to precede and rationally direct *in vivo* drug studies[5]. By experimenting in and benchmarking diverse model systems, we can study a broad range of disease severity, identify useful scenarios for each and explore therapeutic interventions to expedite the end of this pandemic.

4.4 Improving scRNA-seq capture and method dissemination

All the projects described here rely on scRNA-seq as a tool to probe clinical samples and benchmark resulting models. Indeed, RNA sequencing technologies have revolutionized our understanding of cell states and tissue homeostasis[1,2,12,13]. Much progress has been made in increasing cell throughput, resulting in protocols that can isolate and process thousands of cells from a given tissue[14,15]. Seq-Well, Drop-Seq and many other scRNA-seq protocols, rely on bead bound oligos to capture and barcode RNA. mRNA is specifically enriched with a capture bait that targets the polyadenylation region. Barcoding is done at the cell (cell barcode) and transcript (unique molecular identifier, UMI) level. While these plastic capture beads perform their function adequately, innovation in this area has stagnated over the last half decade, and has not kept up with the capture and processing needs of next-generation clinical pipelines. In our research we have improved these RNA capture beads in five distinct ways by 1) developing a quality control pipeline, 2) building from a magnetic scaffold, 3) diversifying capture baits, 4) generally increasing the quantity of RNA captured and barcode diversity and 5) ensure beads are capable of RNA capture after freezing, to ensure compatibility with more user-friendly Seq-Well protocols. Combined, these improvements will boost yield from clinical samples and allow for better throughput when modeling tissues and drug-testing distinct cell states.

To improve bead quality, we established a flow cytometry-based quality control pipeline. By reversibly binding fluorescent oligos tailored to a bead's capture bait we aim to FACS sort and utilize only beads capable of capturing the most mRNA. Additionally, by incorporating photocleavable sequences we can enumerate full length oligos and assess the diversity of the barcodes and UMIs from a given bead batch. Next, experiments building from magnetic scaffolds have shown improvement in bead recovery and retention through the whole Seq-Well protocol. This is critical when working with precious clinical samples. Third, to diversify capture baits, we synthesize a universal sequence capable of binding and extending oligos of interest in concert with poly-A sequences. This greatly broadens the utility of Seq-Well allowing for enrichment of viral transcripts, TCR/BCR sequences, or phenotype defining tumor cell genes at the RNA capture step. Next, by improving the efficiency of our oligo synthesis method we can increase the number of full-length capture sequences on each bead. Lastly, we established a truncated array freezing Seq-Well protocol. This is specifically compatible with our beads and allows us to collect and store

more cells at the tissue dissociation step by delaying downstream processing. This last technique has already been adopted widely in time sensitive or sample limited experiments. Collectively, these improvements have already increased cell, gene, and sample recovery from our Seq-Well experiments and will play a critical role in the adaptation of scRNA-seq pipeline in the clinic.

4.5 Summary

scRNA-seq has already changed our understanding of many tissues and the onset and progress of disease[3,4,16-18]. Technological innovations in single-cell measurements will continue to couple multimodal information with RNA expression (*e.g.* spatial, protein *etc.*) into increasingly powered datasets to further disentangle cellular interactions[19-21]. However, thus far there are still large gaps between academic discovery and clinical translation.

This body of work aims to bridge that gap in several ways: 1) by benchmarking patient samples we can assess *ex vivo* model fidelity and incorporate native extrinsic- and intrinsic-state specific factors to maintain the *in vivo* phenotype. 2) By serially sampling *ex vivo* and animal model systems we can track natural drift in RNA state, infection-induced phenotype changes, or drug response at the system-wide level. Lastly, 3) we have improved accessibility and approachability of these tools by developing quality control benchmarks and ease-of-use improvements to the widely used Seq-Well platform.

While the above work has been critical in applying scRNA-seq technology in novel clinical ways, true patient-facing applications will require further refinement of these models and pipelines. To date, much of our work has focused on model fidelity and disease profiling, with drug efficacy benchmarking a secondary focus. Ideally, as our understanding of disease-induced changes in the RNA state matures and our models become more representative, scRNA-seq experiments on patient derived models will focus more on matching drugs to patients (as genotype directed therapy does now), in well-vetted models of tumor cell state or viral infection[22]. Together we hope this work builds to new personalized medicine pipelines, applicable in PDAC, COVID-19, and beyond.

4.6 References

1. Gierahn, T.M., Wadsworth, M.H., 2nd, Hughes, T.K., Bryson, B.D., Butler, A., Satija, R., Fortune, S., Love, J.C., and Shalek, A.K. (2017). Seq-Well: portable, low-cost RNA sequencing of single cells at high throughput. Nat Methods *14*, 395-398.

2. Hughes, T.K., Wadsworth, M.H., 2nd, Gierahn, T.M., Do, T., Weiss, D., Andrade, P.R., Ma, F., de Andrade Silva, B.J., Shao, S., Tsoi, L.C.*, et al.* (2020). Second-Strand Synthesis-Based Massively Parallel scRNA-Seq Reveals Cellular States and Molecular Features of Human Inflammatory Skin Pathologies. Immunity *53*, 878-894 e877.

3. Raghavan, S., Winter, P. S., Navia, A. W., Williams, H. L., DenAdel, A., Lowder, K. E., Galvez-Reyes, J., Kalekar, R. L., Mulugeta, N., Kapner, K. S., Raghavan, M. S., Borah, A. A., Liu, N., Väyrynen, S. A., Costa, A. D., Ng, R., Wang, J., Hill, E. K., Ragon, D. Y., Brais, L. K., … Shalek, A. K. (2021). Microenvironment drives cell state, plasticity, and drug response in pancreatic cancer. *Cell*, *184*(25), 6119–6137.e26.
4. Ziegler, C., Miao, V. N., Owings, A. H., Navia, A. W., Tang, Y., Bromley, J. D., Lotfy, P., Sloan, M., Laird, H., Williams, H. B., George, M., Drake, R. S., Christian, T., Parker, A., Sindel, C. B., Burger, M. W., Pride, Y., Hasan, M., Abraham, G. E., 3rd, Senitko, M., … Ordovas-Montanes, J. (2021). Impaired local intrinsic immunity to SARS-CoV-2 infection in severe COVID-19. *Cell*, *184*(18), 4713–4733.e22.
5. Bein, A., Kim, S., Goyal, G., Cao, W., Fadel, C., Naziripour, A., Sharma, S., Swenor, B., LoGrande, N., Nurani, A., Miao, V. N., Navia, A. W., Ziegler, C., Montañes, J. O., Prabhala, P., Kim, M. S., Prantil-Baun, R., Rodas, M., Jiang, A., O'Sullivan, L., … Ingber, D. E. (2021). Enteric Coronavirus Infection and Treatment Modeled With an Immunocompetent Human Intestine-On-A-Chip. *Frontiers in pharmacology*, *12*, 718484.
6. Chen, X., & Wood, B. L. (2017). How do we measure MRD in ALL and how should measurements affect decisions. Re: Treatment and prognosis?. *Best practice & research. Clinical haematology*, *30*(3), 237–248.
7. Szczepański, T., Orfão, A., van der Velden, V. H., San Miguel, J. F., & van Dongen, J. J. (2001). Minimal residual disease in leukaemia patients. *The Lancet. Oncology*, *2*(7), 409–417.
8. Jongen-Lavrencic, M., Grob, T., Hanekamp, D., Kavelaars, F. G., Al Hinai, A., Zeilemaker, A., Erpelinck-Verschueren, C., Gradowska, P. L., Meijer, R., Cloos, J., Biemond, B. J., Graux, C., van Marwijk Kooy, M., Manz, M. G., Pabst, T., Passweg, J. R., Havelange, V., Ossenkoppele, G. J., Sanders, M. A., Schuurhuis, G. J., … Valk, P. (2018). Molecular Minimal Residual Disease in Acute Myeloid Leukemia. *The New England journal of medicine*, *378*(13), 1189–1199.
9. Burg, T. P., Godin, M., Knudsen, S. M., Shen, W., Carlson, G., Foster, J. S., Babcock, K., & Manalis, S. R. (2007). Weighing of biomolecules, single cells and single nanoparticles in fluid. *Nature*, *446*(7139), 1066–1069.
10. Olcum, S., Cermak, N., Wasserman, S. C., Christine, K. S., Atsumi, H., Payer, K. R., Shen, W., Lee, J., Belcher, A. M., Bhatia, S. N., & Manalis, S. R. (2014). Weighing nanoparticles in solution

at the attogram scale. *Proceedings of the National Academy of Sciences of the United States of America*, *111*(4), 1310–1315.

11. Hoy S. M. (2014). Ponatinib: a review of its use in adults with chronic myeloid leukaemia or Philadelphia chromosome-positive acute lymphoblastic leukaemia. *Drugs*, *74*(7), 793–806.
12. Picelli, S., Faridani, O. R., Björklund, A. K., Winberg, G., Sagasser, S., & Sandberg, R. (2014). Full-length RNA-seq from single cells using Smart-seq2. *Nature protocols*, *9*(1), 171–181.
13. Tang, F., Barbacioru, C., Wang, Y., Nordman, E., Lee, C., Xu, N., Wang, X., Bodeau, J., Tuch, B. B., Siddiqui, A., Lao, K., & Surani, M. A. (2009). mRNA-Seq whole-transcriptome analysis of a single cell. *Nature methods*, *6*(5), 377–382.
14. Macosko, E. Z., Basu, A., Satija, R., Nemesh, J., Shekhar, K., Goldman, M., Tirosh, I., Bialas, A. R., Kamitaki, N., Martersteck, E. M., Trombetta, J. J., Weitz, D. A., Sanes, J. R., Shalek, A. K., Regev, A., & McCarroll, S. A. (2015). Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets. *Cell*, *161*(5), 1202–1214.
15. Zheng, G. X., Terry, J. M., Belgrader, P., Ryvkin, P., Bent, Z. W., Wilson, R., Ziraldo, S. B., Wheeler, T. D., McDermott, G. P., Zhu, J., Gregory, M. T., Shuga, J., Montesclaros, L., Underwood, J. G., Masquelier, D. A., Nishimura, S. Y., Schnall-Levin, M., Wyatt, P. W., Hindson, C. M., Bharadwaj, R., … Bielas, J. H. (2017). Massively parallel digital transcriptional profiling of single cells. *Nature communications*, *8*, 14049.
16. Song, H., Weinstein, H., Allegakoen, P., Wadsworth, M. H., 2nd, Xie, J., Yang, H., Castro, E. A., Lu, K. L., Stohr, B. A., Feng, F. Y., Carroll, P. R., Wang, B., Cooperberg, M. R., Shalek, A. K., & Huang, F. W. (2022). Single-cell analysis of human primary prostate cancer reveals the heterogeneity of tumor-associated epithelial cell states. *Nature communications*, *13*(1), 141.
17. Hamza, B., Miller, A. B., Meier, L., Stockslager, M., Ng, S. R., King, E. M., Lin, L., DeGouveia, K. L., Mulugeta, N., Calistri, N. L., Strouf, H., Bray, C., Rodriguez, F., Freed-Pastor, W. A., Chin, C. R., Jaramillo, G. C., Burger, M. L., Weinberg, R. A., Shalek, A. K., Jacks, T., … Manalis, S. R. (2021). Measuring kinetics and metastatic propensity of CTCs by blood exchange between mice. *Nature communications*, *12*(1), 5680.
18. Amoozgar, Z., Kloepper, J., Ren, J., Tay, R. E., Kazer, S. W., Kiner, E., Krishnan, S., Posada, J. M., Ghosh, M., Mamessier, E., Wong, C., Ferraro, G. B., Batista, A., Wang, N., Badeaux, M., Roberge, S., Xu, L., Huang, P., Shalek, A. K., Fukumura, D., … Jain, R. K. (2021). Targeting Treg cells with GITR activation alleviates resistance to immunotherapy in murine glioblastomas. *Nature communications*, *12*(1), 2582.

19. Stickels, R. R., Murray, E., Kumar, P., Li, J., Marshall, J. L., Di Bella, D. J., Arlotta, P., Macosko, E. Z., & Chen, F. (2021). Highly sensitive spatial transcriptomics at near-cellular resolution with Slide-seqV2. *Nature biotechnology*, *39*(3), 313–319.
20. Lee, J. H., Daugharthy, E. R., Scheiman, J., Kalhor, R., Yang, J. L., Ferrante, T. C., Terry, R., Jeanty, S. S., Li, C., Amamoto, R., Peters, D. T., Turczyk, B. M., Marblestone, A. H., Inverso, S. A., Bernard, A., Mali, P., Rios, X., Aach, J., & Church, G. M. (2014). Highly multiplexed subcellular RNA sequencing in situ. *Science (New York, N.Y.)*, *343*(6177), 1360–1363.
21. Moffitt, J. R., Hao, J., Wang, G., Chen, K. H., Babcock, H. P., & Zhuang, X. (2016). High-throughput single-cell gene-expression profiling with multiplexed error-robust fluorescence in situ hybridization. *Proceedings of the National Academy of Sciences of the United States of America*, *113*(39), 11046–11051.
22. Bernards R. (2012). A missing link in genotype-directed cancer therapy. *Cell*, *151*(3), 465–468.

Appendix A: Impaired local intrinsic immunity to SARS-CoV-2 infection in severe COVID-19

AUTHORS

Carly G. K. Ziegler*, Vincent N. Miao*, Anna H. Owings*, Andrew W. Navia*, Ying Tang*, Joshua D. Bromley*, Peter Lotfy, Meredith Sloan, Hannah Laird, Haley B. Williams, Micayla George, Riley Drake, Taylor Christian, Adam Parker, L. Campbell Behlen, Molly W. Burger, Yilianys Pride, Kenneth J. Wilson, Mohammad Hasan, George E. Abraham, Michal Senitko, Tanya O. Robinson, Alex K. Shalek#, Bruce H. Horwitz#, Sarah C. Glover#, Jose Ordovas-Montanes#

* these authors contributed equally

# these senior authors contributed equally

Supplementary Table S5.1. Cell Type Marker Genes

Due to its size, this table will be made available upon request. Related to Figures 1, 2, Supplementary Figure 3

Supplementary Table S5.2. Differentially Expressed Genes Between Cell Types from Control WHO 0 vs. COVID-19 WHO 1-5 (mild/moderate)

Due to its size, this table will be made available upon request. Related to Figure 3

Supplementary Table S5.3. Differentially Expressed Genes Between Cell Types from Control WHO 0 vs. COVID-19 WHO 6-8 (severe)

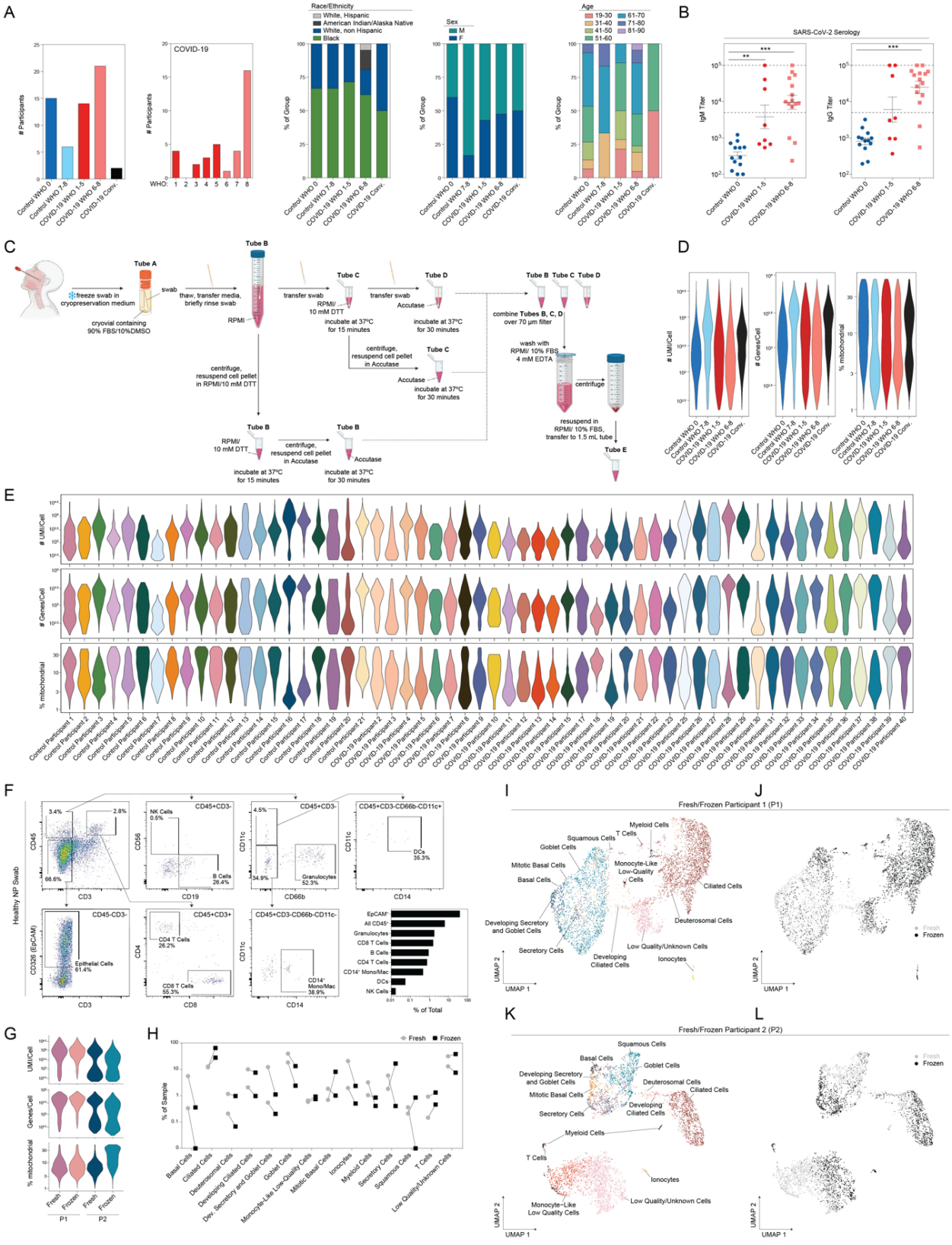
Due to its size, this table will be made available upon request. Related to Figure 3

Supplementary Table S5.4. Differentially Expressed Genes Between Cell Types from COVID-19 WHO 1-5 (mild/moderate) vs. COVID-19 WHO 6-8 (severe)

Due to its size, this table will be made available upon request. Related to Figure 3

Supplementary Table S5.5. Common Differentially Expressed Genes between SARS-CoV-2 RNA+ cells and Bystander Cells

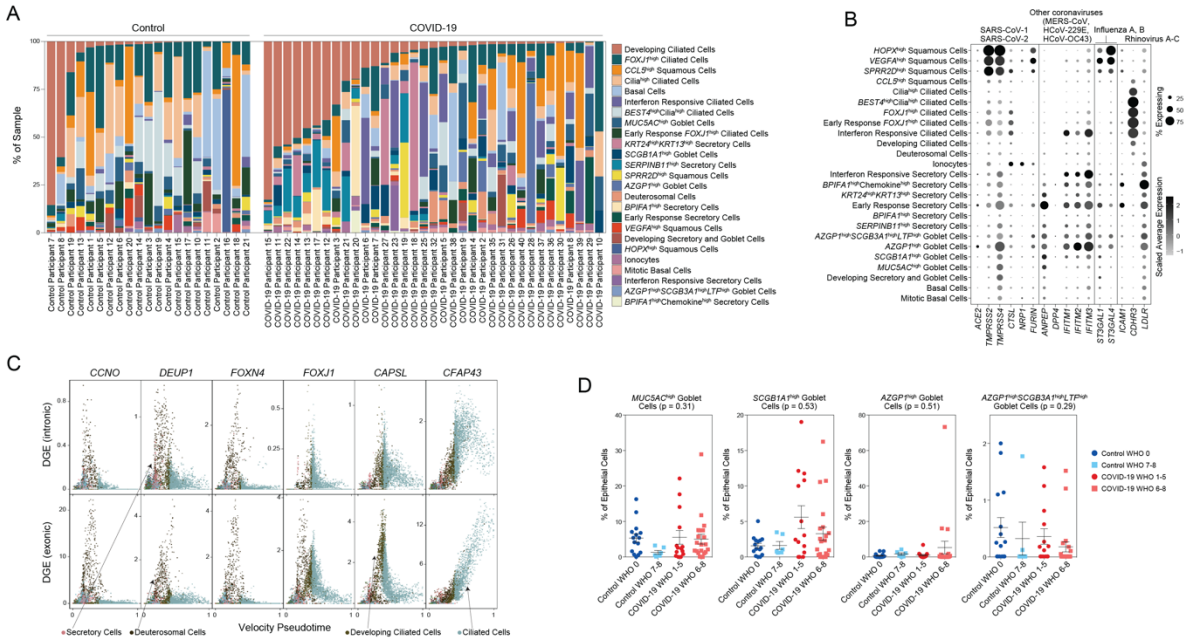
Due to its size, this table will be made available upon request. Related to Figure 6


A
Participants
Control WHO 0
Control WHO 7-8
COVID-19 WHO 1-5
COVID-19 WHO 6-8
COVID-19 Conv.
COVID-19
WHO: 1 2 3 4 5 6 7 8
Race/Ethnicity
White, Hispanic
American Indian/Alaska Native
White, non Hispanic
Black
% of Group
Sex
M
F
Age
19-30
31-40
41-50
51-60
61-70
71-80
81-90
B
SARS-CoV-2 Serology
IgM Titer
IgG Titer
Control WHO 0
COVID-19 WHO 1-5
COVID-19 WHO 6-8
C
Tube A
swab
freeze swab in cryopreservation medium
cryovial containing 90% FBS/10%DMSO
thaw, transfer media, briefly rinse swab
Tube B
RPMI
transfer swab
Tube C
RPMI/ 10 mM DTT
incubate at 37°C for 15 minutes
transfer swab
Tube D
Accutase
incubate at 37°C for 30 minutes
centrifuge, resuspend cell pellet in RPMI/10 mM DTT
centrifuge, resuspend cell pellet in Accutase
Tube B Tube C Tube D
combine Tubes B, C, D over 70 µm filter
wash with RPMI/ 10% FBS 4 mM EDTA
centrifuge
resuspend in RPMI/ 10% FBS, transfer to 1.5 mL tube
Tube E
D
UMI/Cell
Genes/Cell
% mitochondrial
E
UMI/Cell
Genes/Cell
% mitochondrial
Control Participant 1
Control Participant 2
Control Participant 3
Control Participant 4
Control Participant 5
Control Participant 6
Control Participant 7
Control Participant 8
Control Participant 9
Control Participant 10
Control Participant 11
Control Participant 12
Control Participant 13
Control Participant 14
Control Participant 15
Control Participant 16
Control Participant 17
Control Participant 18
Control Participant 19
Control Participant 20
COVID-19 Participant 21
COVID-19 Participant 2
COVID-19 Participant 3
COVID-19 Participant 4
COVID-19 Participant 5
COVID-19 Participant 6
COVID-19 Participant 7
COVID-19 Participant 8
COVID-19 Participant 9
COVID-19 Participant 10
COVID-19 Participant 11
COVID-19 Participant 12
COVID-19 Participant 13
COVID-19 Participant 14
COVID-19 Participant 15
COVID-19 Participant 16
COVID-19 Participant 17
COVID-19 Participant 18
COVID-19 Participant 19
COVID-19 Participant 20
COVID-19 Participant 21
COVID-19 Participant 22
COVID-19 Participant 23
COVID-19 Participant 25
COVID-19 Participant 26
COVID-19 Participant 27
COVID-19 Participant 28
COVID-19 Participant 29
COVID-19 Participant 30
COVID-19 Participant 31
COVID-19 Participant 32
COVID-19 Participant 33
COVID-19 Participant 34
COVID-19 Participant 35
COVID-19 Participant 36
COVID-19 Participant 37
COVID-19 Participant 38
COVID-19 Participant 39
COVID-19 Participant 40
F
Healthy NP Swab
CD45
CD3
CD56
CD19
NK Cells
B Cells
CD11c
CD66b
Granulocytes
CD14
DCs
CD326 (EpCAM)
Epithelial Cells
CD4
CD8
CD4 T Cells
CD8 T Cells
EpCAM+
All CD45-
CD14+ Mono/Mac
NK Cells
% of Total
G
UMI/Cell
Genes/Cell
% mitochondrial
Fresh
Frozen
P1
P2
H
Fresh
Frozen
% of Sample
Basal Cells
Ciliated Cells
Deuterosomal Cells
Developing Ciliated Cells
Dev. Secretory and Goblet Cells
Goblet Cells
Monocyte-Like Low-Quality Cells
Mitotic Basal Cells
Ionocytes
Myeloid Cells
Secretory Cells
Squamous Cells
T Cells
Low Quality/Unknown Cells
I
Squamous Cells
Myeloid Cells
T Cells
Goblet Cells
Mitotic Basal Cells
Basal Cells
Monocyte-Like Low-Quality Cells
Ciliated Cells
Developing Secretory and Goblet Cells
Deuterosomal Cells
Secretory Cells
Low Quality/Unknown Cells
Developing Ciliated Cells
Ionocytes
UMAP 1
UMAP 2
J
Fresh/Frozen Participant 1 (P1)
Fresh
Frozen
K
Squamous Cells
Basal Cells
Goblet Cells
Developing Secretory and Goblet Cells
Deuterosomal Cells
Ciliated Cells
Mitotic Basal Cells
Developing Ciliated Cells
Secretory Cells
Myeloid Cells
T Cells
Ionocytes
Monocyte-Like Low Quality Cells
Low Quality/Unknown Cells
L
Fresh/Frozen Participant 2 (P2)
Fresh
Frozen

Supplementary Figure 1. Cohort and Cellular Composition of Nasopharyngeal Swabs

Related to Figure 1, Table 1

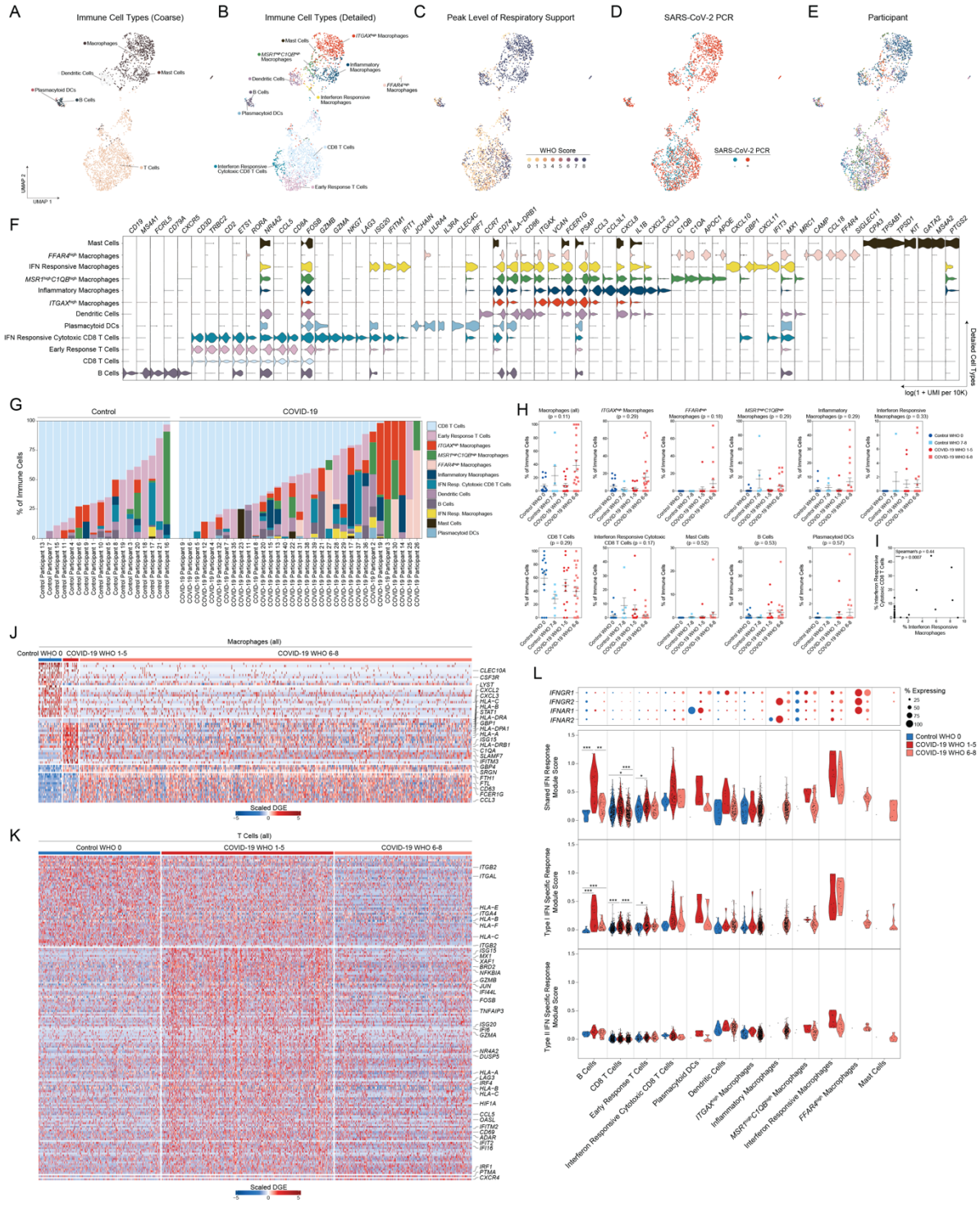
(A) Cohort composition and participant demographics (see also **Table 1**). **(B)** SARS-CoV-2 serology: IgM (left) and IgG (right) titers from a subset of Control WHO 0 (blue circles, n=13) and COVID-19 (red circles, mild/moderate: n=8; pink squares, severe: n=15) participants. Plasma samples taken on same day of nasopharyngeal swab. Statistical testing by Kruskal-Wallis test with Dunn's post hoc testing. Asterisks represent results from Dunn's test: ** $p < 0.01$, *** $p < 0.001$. Dashed lines: lower limit of detection: 100; upper limit of detection: 100,000; positive threshold: 5,000. **(C)** Detailed schematic of sample preparation and cell processing from nasal swabs (created with BioRender). **(D)** Single-cell quality metrics by group (after filtering for low-quality cells, see **Methods**). **(E)** Single-cell quality metrics by participant (after filtering for low quality cells). **(F)** Flow cytometry and gating scheme of cells from a representative fresh nasopharyngeal swab from a healthy participant. Bottom right: quantification of cellular proportions. **(G)** Quality metrics for matched fresh vs. frozen nasal swabs from two healthy participants (P1 and P2). **(H)** Percent composition of each cell type by processing type: fresh (grey circles) or frozen (black squares). **(I)** UMAP of cell types from P1. **(J)** UMAP from P1 as in **I**, colored by fresh (grey) vs. frozen (black). **(K)** UMAP of cell types from P2. **(L)** UMAP from P2 as in **K**, colored by fresh (grey) vs. frozen (black).



Supplementary Figure S5.2. COVID-19-induced changes to epithelial diversity and differentiation

Related to Figure 5.2

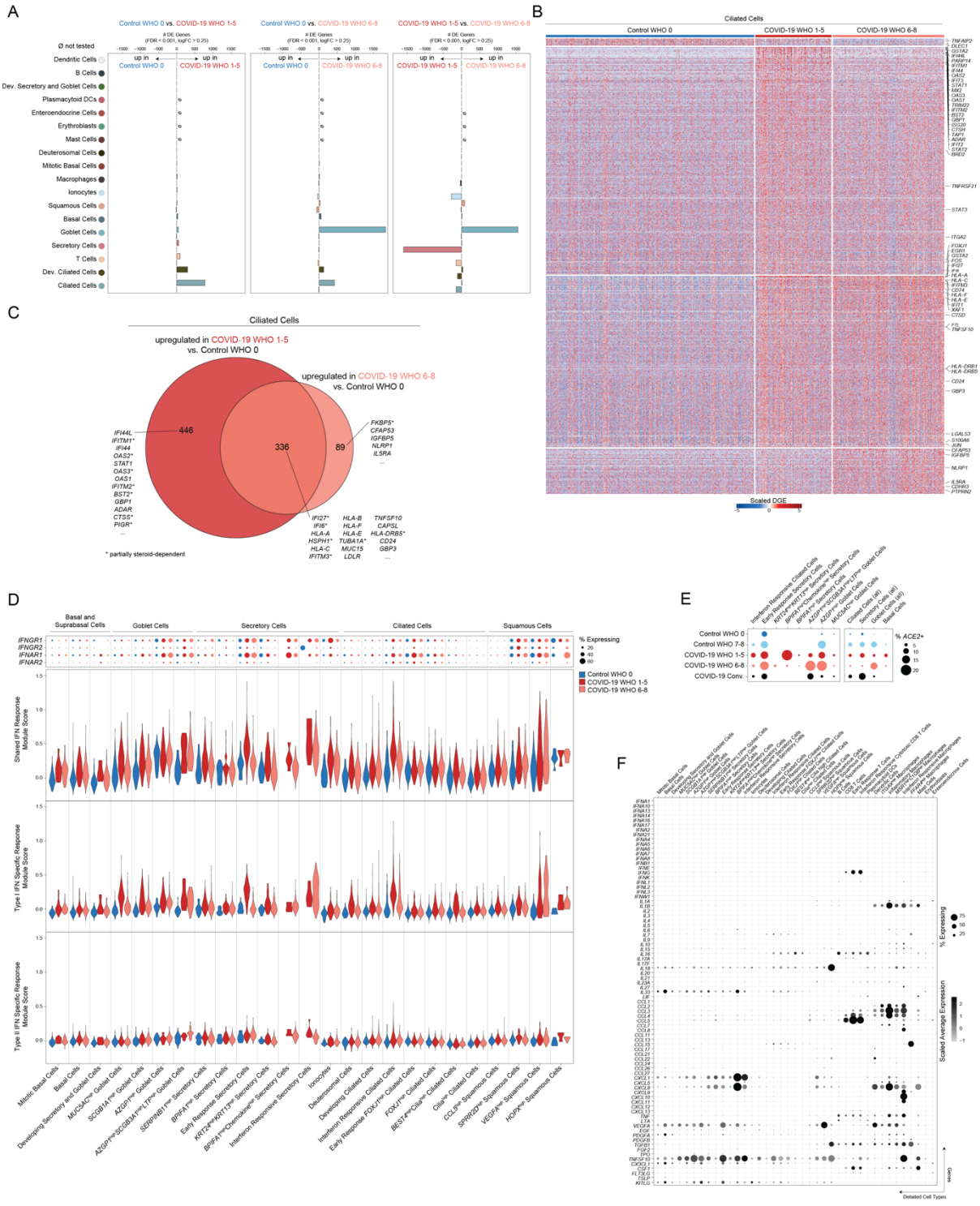
(A) Proportional abundance of detailed epithelial cell types by participant. **(B)** Expression of entry factors for SARS-CoV-2 and other common upper respiratory viruses among detailed epithelial cell types. Dot size represents fraction of cell type (rows) expressing a given gene (columns). Dot hue represents average expression. **(C)** Plot of gene expression by epithelial cell velocity pseudotime. Select genes significantly associated with ciliated cell pseudotime. Points colored by coarse cell type annotations. Top: alignment to unspliced (intronic) regions. Bottom: alignment to spliced (exonic) regions. **(D)** Proportion of Goblet Cell subtypes (detailed annotation) by sample, normalized to all epithelial cells. Statistical test above graph represents Kruskal-Wallis test results across all cohorts (following Bonferroni-correction).



Supplementary Figure S5.3. COVID-19-induced changes to nasopharynx-resident immune cells

Related to Figures 5.1 and 5.3

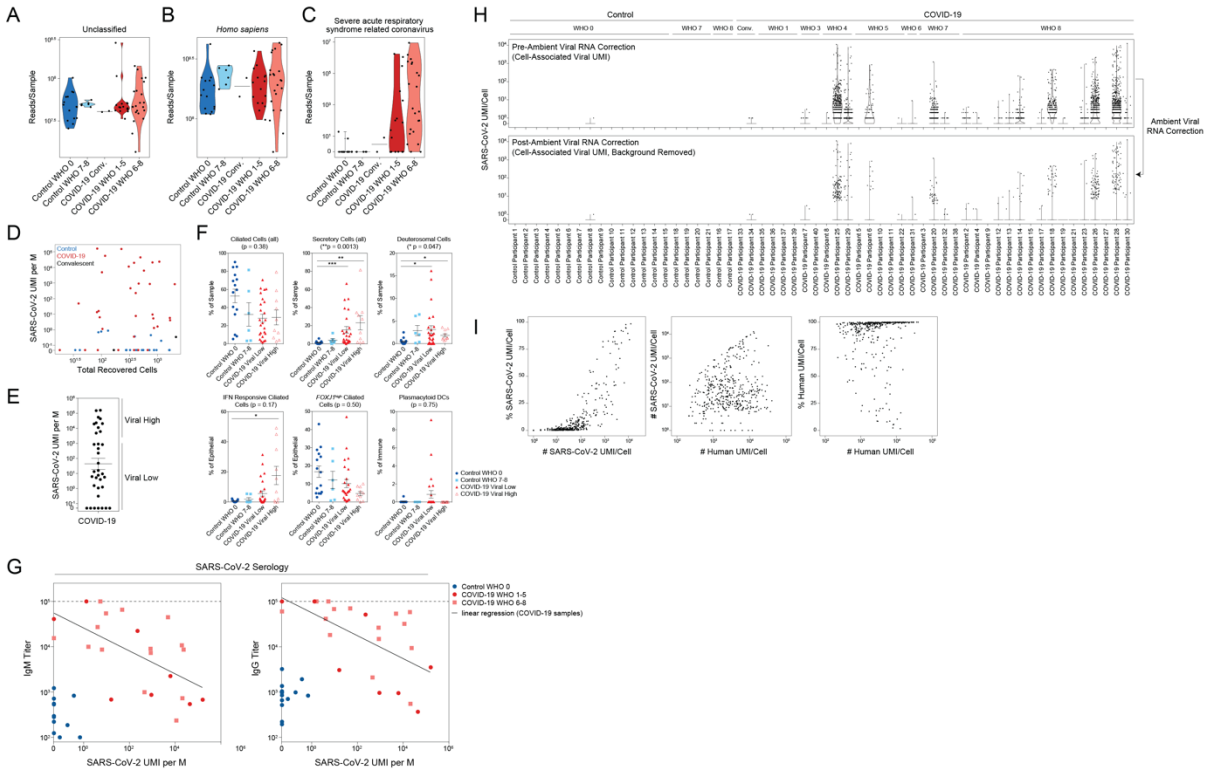
(A) UMAP of 3,640 immune cells following re-clustering, colored by coarse cell types. **(B)** UMAP as in **A**, colored by detailed cell annotations. **(C)** UMAP as in **A**, colored by level of respiratory support (WHO illness severity scale). **(D)** UMAP as in **A**, colored by SARS-CoV-2 PCR status at time of swab. **(E)** UMAP as in **A**, colored by participant. **(F)** Violin plots of cluster marker genes (FDR < 0.01) for detailed immune cell type annotations (as in **B**). **(G)** Proportional abundance of detailed immune cell types by participant. **(H)** Proportion of immune cell subtypes by sample and cohort, normalized to all immune cells. Statistical test above graph represents Kruskal-Wallis test results across all cohorts (following Bonferroni-correction). **(I)** Heatmap of significantly DE genes between Macrophages (all, coarse annotation) from different disease cohorts. **(J)** Heatmap of significantly DE genes between T Cells (all, coarse annotation) from different disease cohorts. **(K)** Top: Dot plot of *IFNGR1/2* and *IFNAR1/2* gene expression among all detailed immune subtypes. Bottom: Violin plots of gene module scores, split by Control WHO 0 (blue), COVID-19 WHO 1-5 (red), and COVID-19 WHO 6-8 (pink). Gene modules represent transcriptional responses of human basal cells from the nasal epithelium following *in vitro* treatment with IFNA or IFNG. Significance by Wilcoxon signed-rank test. P-values following Bonferroni-correction: * $p< 0.05$, ** $p < 0.01$, *** $p < 0.001$.


A
Control WHO 0 vs. COVID-19 WHO 1-5
Control WHO 0 vs. COVID-19 WHO 6-8
COVID-19 WHO 1-5 vs. COVID-19 WHO 6-8
DE Genes
(FDR < 0.001, logFC > 0.25)
up in
Control WHO 0
COVID-19 WHO 1-5
COVID-19 WHO 6-8
Ø not tested
Dendritic Cells
B Cells
Dev. Secretory and Goblet Cells
Plasmacytoid DCs
Enteroendocrine Cells
Erythroblasts
Mast Cells
Deuterosomal Cells
Mitotic Basal Cells
Macrophages
Ionocytes
Squamous Cells
Basal Cells
Goblet Cells
Secretory Cells
T Cells
Dev. Ciliated Cells
Ciliated Cells
B
Ciliated Cells
Control WHO 0
COVID-19 WHO 1-5
COVID-19 WHO 6-8
Scaled DGE
C
Ciliated Cells
upregulated in COVID-19 WHO 1-5
vs. Control WHO 0
upregulated in COVID-19 WHO 6-8
vs. Control WHO 0
446
336
89
IFI44L
IFITM1*
IFI44
OAS2*
STAT1
OAS3*
OAS1
IFITM2*
BST2*
GBP1
ADAR
CTSS*
PIGR*
...
IFI27*
IFI6*
HLA-A
HSPH1*
HLA-C
IFITM3*
HLA-B
HLA-F
HLA-E
TUBA1A*
MUC15
LDLR
TNFSF10
CAPSL
HLA-DRB5*
CD24
GBP3
...
FKBP5*
CFAP53
IGFBP5
NLRP1
IL5RA
...
* partially steroid-dependent
D
Basal and Suprabasal Cells
Goblet Cells
Secretory Cells
Ciliated Cells
Squamous Cells
IFNGR1
IFNGR2
IFNAR1
IFNAR2
% Expressing
Control WHO 0
COVID-19 WHO 1-5
COVID-19 WHO 6-8
Shared IFN Response Module Score
Type I IFN Specific Response Module Score
Type II IFN Specific Response Module Score
Mitotic Basal Cells
Basal Cells
Developing Secretory and Goblet Cells
MUC5AC^hi^ Goblet Cells
SCGB1A1^hi^ Goblet Cells
AZGP1^hi^ Goblet Cells
AZGP1^hi^SCGB3A1^hi^LTF^hi^ Goblet Cells
SERPINB11^hi^ Goblet Cells
BPIFA1^hi^ Secretory Cells
Early Response Secretory Cells
KRT24^hi^KRT13^hi^ Secretory Cells
BPIFA1^hi^Chemokine^hi^ Secretory Cells
Interferon Responsive Secretory Cells
Ionocytes*
Deuterosomal Cells
Developing Ciliated Cells
Interferon Responsive Ciliated Cells
Early Response FOXJ1^low^ Ciliated Cells
FOXJ1^hi^ Ciliated Cells
BEST4^hi^Cilia^hi^ Ciliated Cells
Cilia^hi^ Ciliated Cells
CCL5^hi^ Squamous Cells
SPRR2D^hi^ Squamous Cells
VEGFA^hi^ Squamous Cells
HOPX^hi^ Squamous Cells
E
Interferon Responsive Ciliated Cells
Early Response Secretory Cells
KRT24^hi^KRT13^hi^ Secretory Cells
BPIFA1^hi^Chemokine^hi^ Secretory Cells
AZGP1^hi^ Secretory Cells
AZGP1^hi^SCGB3A1^hi^LTF^hi^ Goblet Cells
MUC5AC^hi^ Goblet Cells
Ciliated Cells (all)
Secretory Cells (all)
Goblet Cells (all)
Basal Cells
Control WHO 0
Control WHO 7-8
COVID-19 WHO 1-5
COVID-19 WHO 6-8
COVID-19 Conv.
% ACE2+
F
% Expressing
Scaled Average Expression
Genes
Detailed Cell Types

Supplementary Figure S5.4. Cell-type specific and shared transcriptional responses to SARS-CoV-2 infection

Related to Figure 5.3

(A) Abundance of significant differentially expressed genes by coarse cell type between Control WHO 0 and COVID-19 WHO 1-5 samples (left), Control WHO 0 and COVID-19 WHO 6-8 samples (middle) and COVID-19 WHO 1-5 vs. COVID-19 WHO 6-8 samples (right). FDR-corrected $p < 0.001$, log2 fold change > 0.25. **(B)** Heatmap of significantly DE genes between Ciliated Cells (all, coarse annotation) from different disease cohorts. **(C)** Venn diagram of significantly upregulated genes among Ciliated Cells between COVID-19 WHO 1-5 vs Control WHO 0 (red) and COVID-19 WHO 6-8 vs. Control WHO 0 (pink). Asterisk: genes impacted by steroid treatment within each cohort. **(D)** Interferon gene module scores across all detailed epithelial cell types, split by Control WHO 0 (blue), COVID-19 WHO 1-5 (red), and COVID-19 WHO 6-8 (pink). Gene modules represent transcriptional responses of human basal cells from the nasal epithelium following *in vitro* treatment with IFNA or IFNG. **(E)** Dot plot of *ACE2* expression across select coarse and detailed epithelial cell types and subsets. **(F)** Dot plot of interferon and cytokine expression among detailed epithelial and immune cell types.

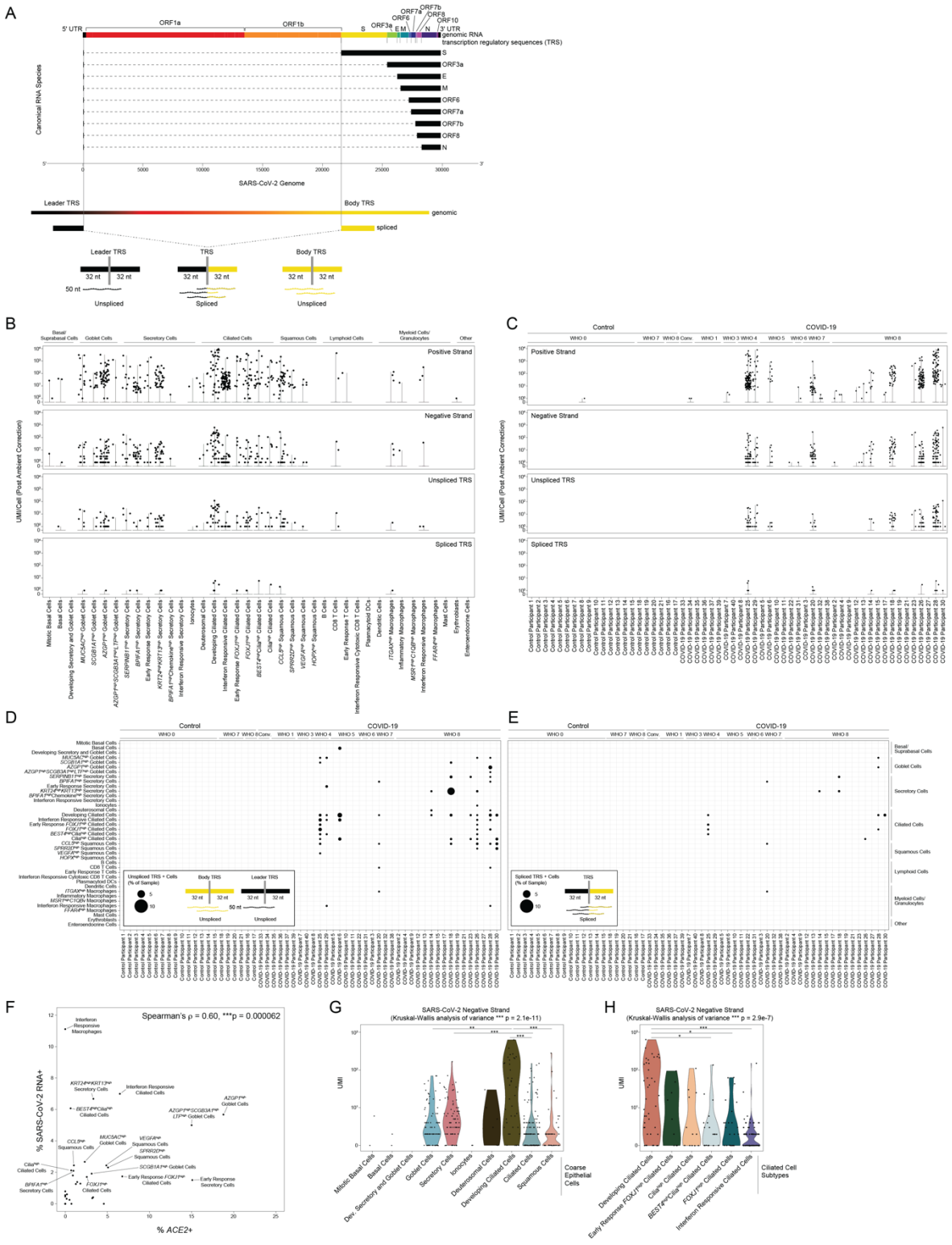


Supplementary Figure S5.5. Detection of SARS-CoV-2 RNA from single-cell RNA-seq data

Related to Figures 5.4 and 5.5

(A) Metatranscriptomic classification of all single-cell RNA-seq reads using Kraken2: reads per sample annotated as unclassified. **(B)** Metatranscriptomic classification of all single-cell RNA-seq reads using Kraken2: reads per sample annotated as *Homo sapiens.* **(C)** Metatranscriptomic classification of all single-cell RNA-seq reads using Kraken2: reads per sample annotated as SARS-related coronaviruses. **(D)** Total recovered cells per sample vs. normalized abundance of SARS-CoV-2 aligning UMI from all single-cell RNA-seq reads (including those derived from ambient/low-quality cell barcodes). **(E)** Normalized abundance of SARS-CoV-2 aligning UMI from all single-cell RNA-seq reads across all COVID-19 participants. Dashed line represents partition between "Viral High" vs "Viral Low" samples. **(F)** Proportional abundance of selected cell types according to total SARS-CoV-2 abundance among COVID-19 samples. Statistical test above graph represents Kruskal-Wallis test statistic across all cohorts. Statistical significance asterisks within box represent significant results from Dunn's post-hoc testing. Bonferroni-corrected p-value: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. **(G)** Abundance of SARS-CoV-2 aligning UMI/cell by participant prior to (top) and following (bottom) ambient viral RNA correction. **(H)** Quality metrics among 415 SARS-CoV-2 RNA+ cells (associated with high-quality cell barcodes and following ambient viral RNA correction). Left: abundance of SARS-CoV-2 aligning UMI vs. percent of all aligned reads (per cell barcode) aligning to SARS-CoV-2. Middle: abundance of human (GRCh38)-aligning UMI

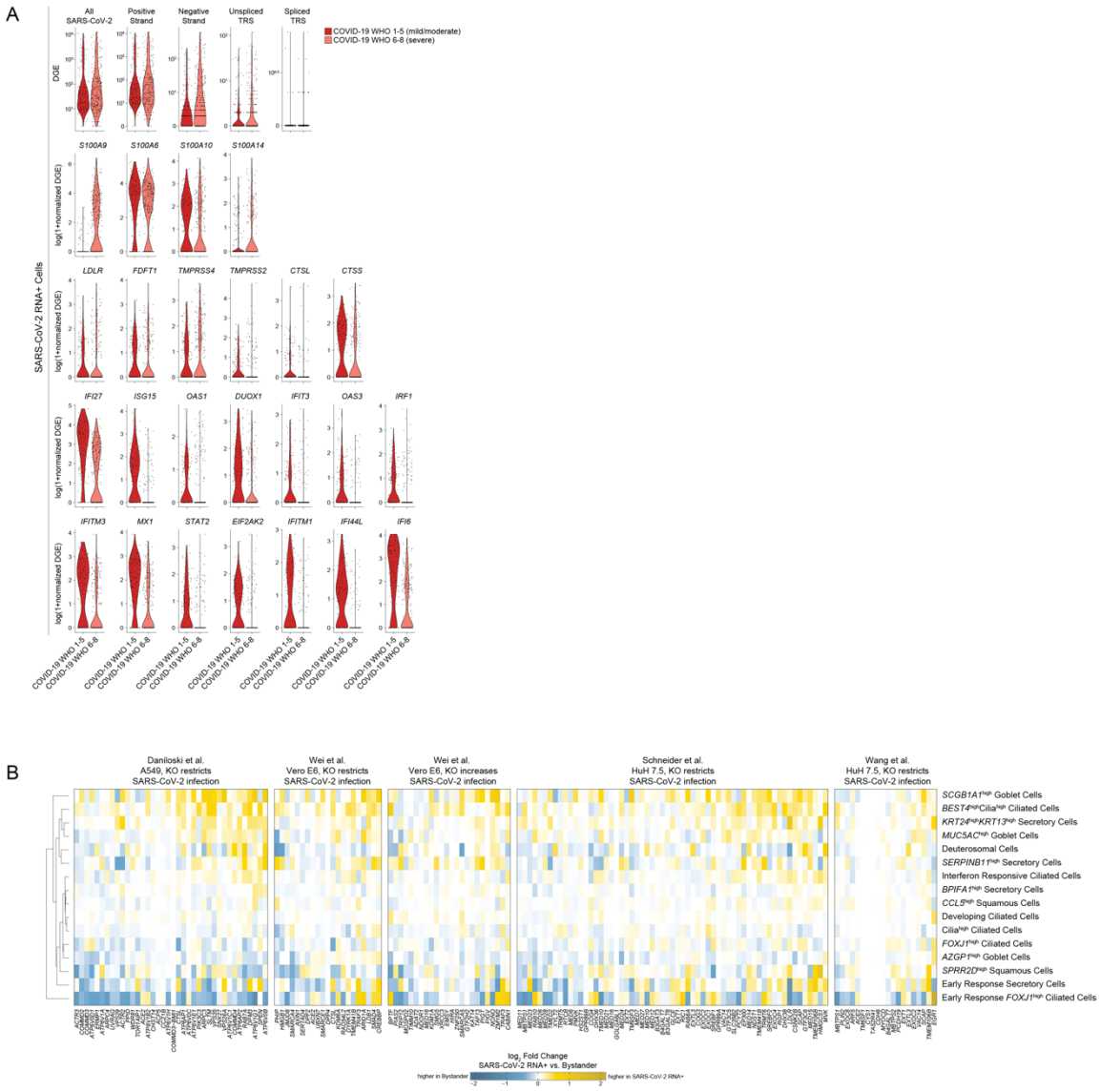
vs. abundance of SARS-CoV-2 aligning UMI. Right: abundance of human (GRCh38) aligning UMI vs. percent of all aligned reads (per cell barcode) aligning to human genes.



Supplementary Figure S5.6. SARS-CoV-2 RNA species and cell types containing viral reads

Related to Figures 5.4 and 5.5

(A) Schematic of method to distinguish unspliced from spliced SARS-CoV-2 RNA species by searching for reads which align across a spliced or genomic Transcription Regulatory Sequence (TRS). **(B)** Abundance of SARS-CoV-2 aligning UMI/Cell per detailed cell type (following ambient viral RNA correction), split by UMI aligning to the viral positive strand, negative strand, 70-mer region across an unspliced TRS, and 70-mer region across a spliced TRS. **(C)** Abundance of SARS-CoV-2 aligning UMI/Cell per participant (following ambient viral RNA correction), split by UMI aligning to the viral positive strand, negative strand, 70-mer region across an unspliced TRS, and 70-mer region across a spliced TRS. **(D)** Dot plot of SARS-CoV-2 unspliced TRS aligning UMI by participant (columns) and detailed cell type (rows). **(E)** Dot plot of SARS-CoV-2 spliced TRS aligning UMI by participant (columns) and detailed cell type (rows). **(F)** Percent ACE2+ cells vs. percent SARS-CoV-2 RNA+ (after ambient correction) by detailed cell type. Including only cells from COVID-19 participants. Statistical testing using spearman's correlation. **(G)** Abundance of SARS-CoV-2 negative strand aligning reads by coarse epithelial cell types. Statistical significance by Kruskal-Wallis test (p-value outside box). Asterisks within box: pairwise wilcox post test, Bonferroni-corrected: *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$. **(H)** Abundance of SARS-CoV-2 negative strand aligning reads by detailed Ciliated Cell subtypes. Statistical significance by Kruskal-Wallis test (p-value outside box). Asterisks within box: pairwise wilcox post test, Bonferroni-corrected: *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$



Supplementary Figure S5.7. **Intrinsic and bystander responses to SARS-CoV-2 infection**

Related to Figure 5.6

(A) Violin plots of select genes upregulated in SARS-CoV-2 RNA+ Cells when compared to matched bystanders. Plotting only SARS-CoV-2 RNA+ Cells from COVID-19 WHO 1-5 participants (red) and COVID-19 WHO 6-8 participants (pink). Top row: SARS-CoV-2 RNA expression by alignment type. **(B)** Heatmaps of log fold changes between SARS-CoV-2 RNA+ cells and bystander cells by cell types. Gene sets derived from four CRISPR screens for important host factors in the SARS-CoV-2 viral life cycle. Restricted to cell types with at least 5 SARS-CoV-2 RNA+ cells. Yellow: upregulated among SARS-CoV-2 RNA+ cells, blue: upregulated among bystander cells.

Appendix B: The tumor microenvironment drives transcriptional phenotypes and their plasticity in metastatic pancreatic cancer

Srivatsan Raghavan*, Peter S. Winter *, Andrew W. Navia*, Hannah L. Williams*, Alan DenAdel, Radha L. Kalekar, Jennyfer Galvez-Reyes, Kristen E. Lowder, Nolawit Mulugeta, Manisha S. Raghavan, Ashir A. Borah, Kevin S. Kapner, Sara A. Väyrynen, Andressa Dias Costa, Raymond W.S. Ng, Junning Wang, Emma Reilly, Dorisanne Y. Ragon, Lauren K. Brais, Alex M. Jaeger, Liam F. Spurr, Yvonne Y. Li, Andrew D. Cherniack , Isaac Wakiro, Asaf Rotem , Bruce E. Johnson , James M. McFarland, Ewa T. Sicinska, Tyler E. Jacks, Thomas E. Clancy, Kimberly Perez, Douglas A. Rubinson, Kimmie Ng, James M. Cleary, Lorin Crawford, Scott R. Manalis, Jonathan A. Nowak, Brian M. Wolpin#, William C. Hahn#, Andrew J. Aguirre#, Alex K. Shalek#

*These authors contributed equally to this work
#These authors contributed equally to this work

Supplemental Table S2.1. Cohort patient characteristics.
Due to its size, this table will be made available upon request. Related to Figure 1

Supplemental Table S2.2. Normal cell type markers.
Due to its size, this table will be made available upon request. Related to Figures 1, 5 & 6

Supplemental Table S2.3. Malignant phenotype single-cell gene correlates.
Due to its size, this table will be made available upon request. Related to Figure 2

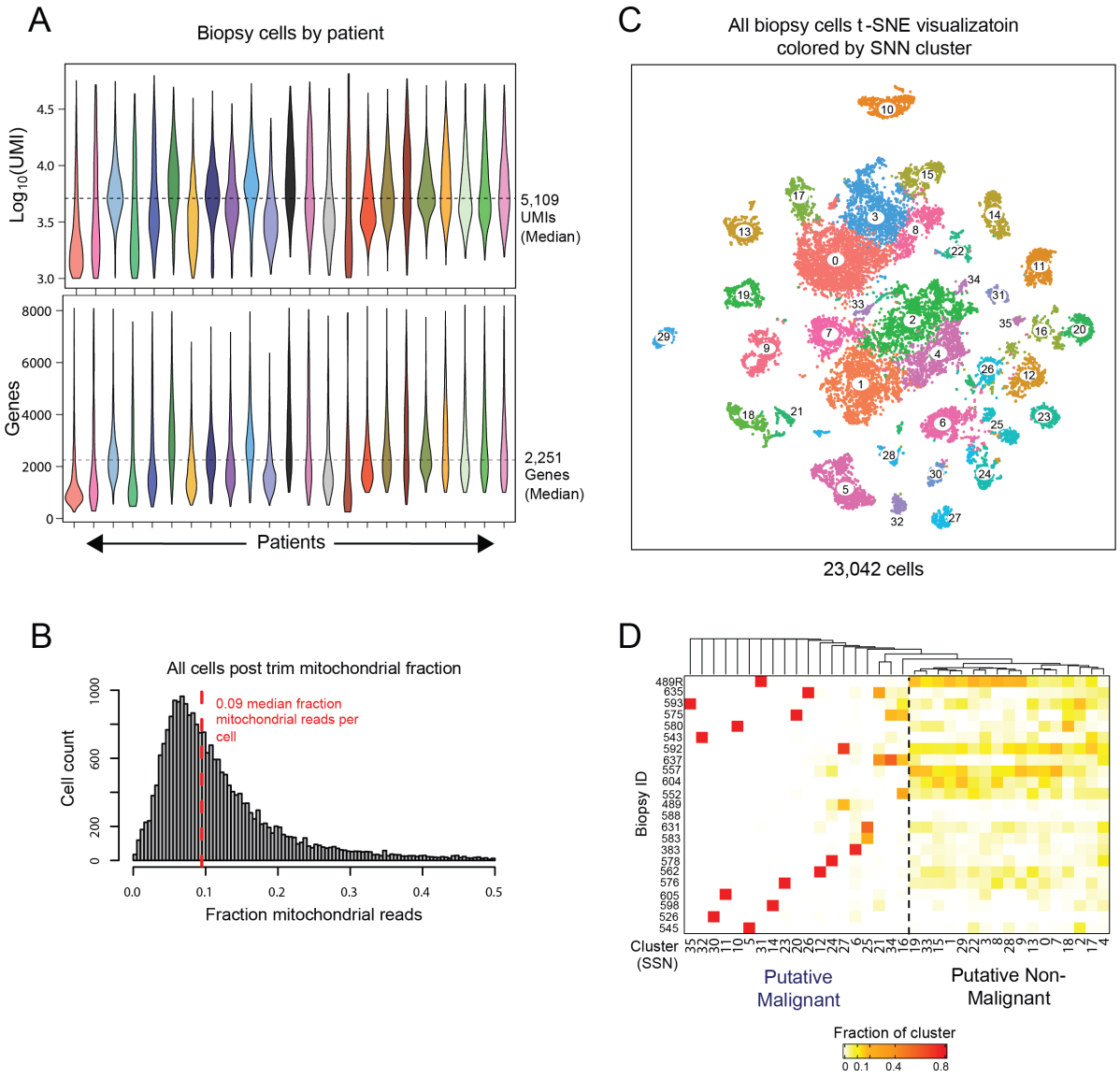
Supplemental Table S2.4. mIF marker combinations and cell counts.
Due to its size, this table will be made available upon request. Related to Figure 2

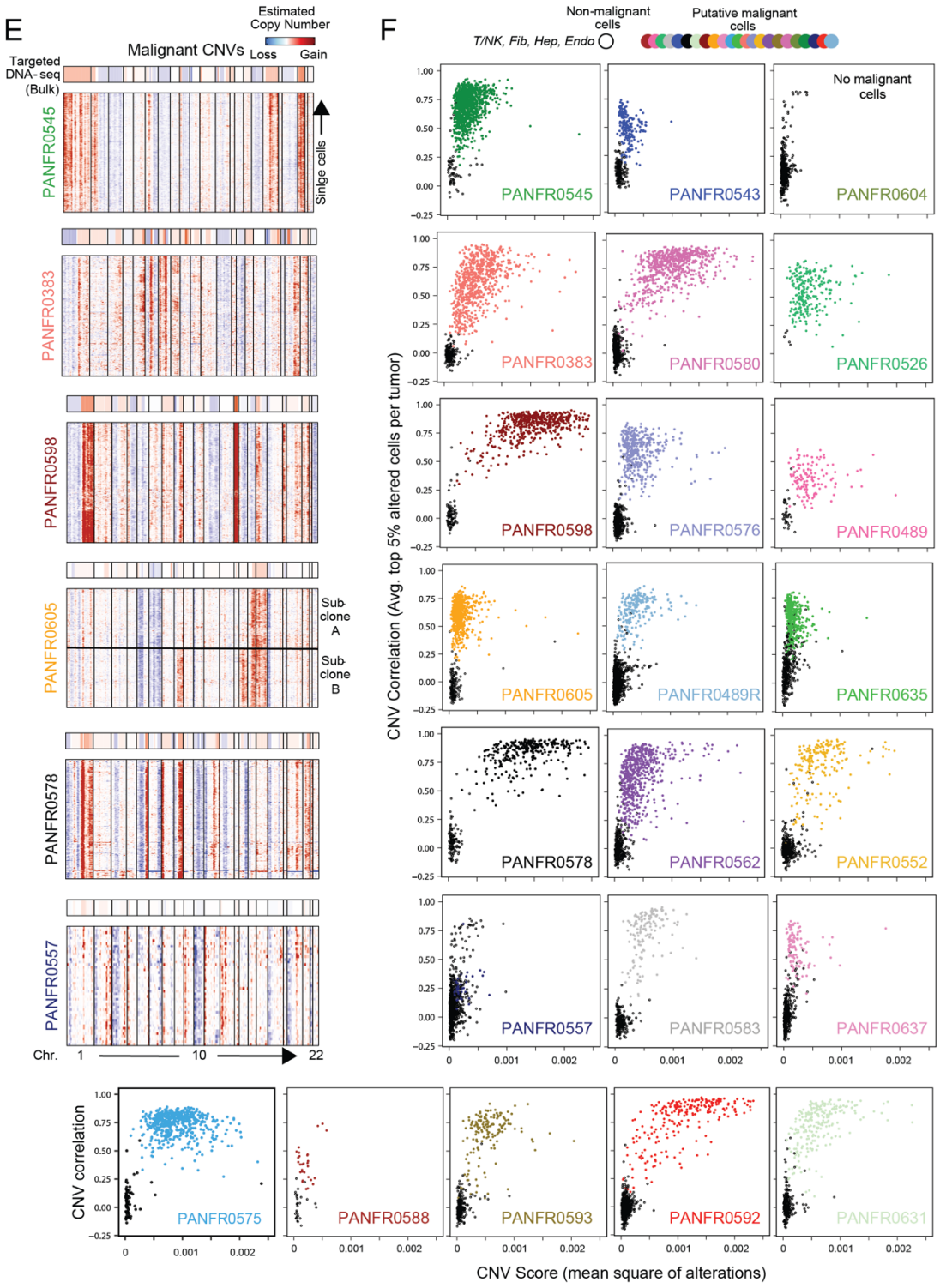
Supplemental Table S2.5. Organoid- and in vivo malignant-specific gene expression features.
Due to its size, this table will be made available upon request. Related to Figure 4

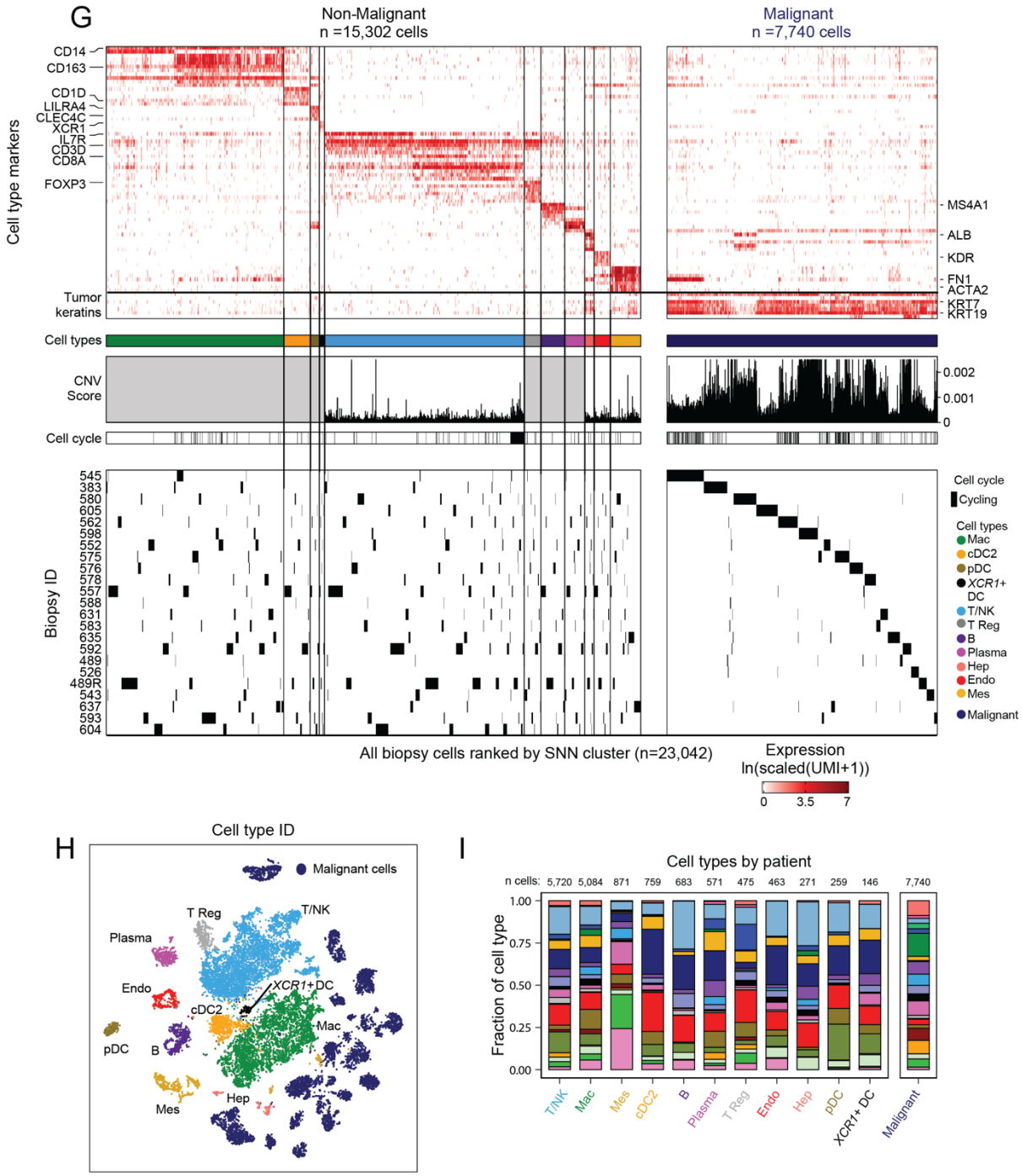
Supplemental Table S2.6. Organoid and cell line models and media formulations for perturbation experiments.

Due to its size, this table will be made available upon request. Related to Figure 7

Supplemental Table S2.7. Subtype-specific autocrine and paracrine secreted factors.
Due to its size, this table will be made available upon request. Related to Figure 7


A
Biopsy cells by patient
Log10(UMI)
4.5
4.0
3.5
3.0
5,109
UMIs
(Median)
Genes
8000
6000
4000
2000
0
2,251
Genes
(Median)
Patients
C
All biopsy cells t-SNE visualizatoin
colored by SNN cluster
23,042 cells
B
All cells post trim mitochondrial fraction
0.09 median fraction
mitochondrial reads per
cell
Cell count
1000
600
200
0
0.0
0.1
0.2
0.3
0.4
0.5
Fraction mitochondrial reads
D
Biopsy ID
Cluster
(SSN)
Putative
Malignant
Putative Non-
Malignant
Fraction of cluster
0 0.1 0.4 0.8

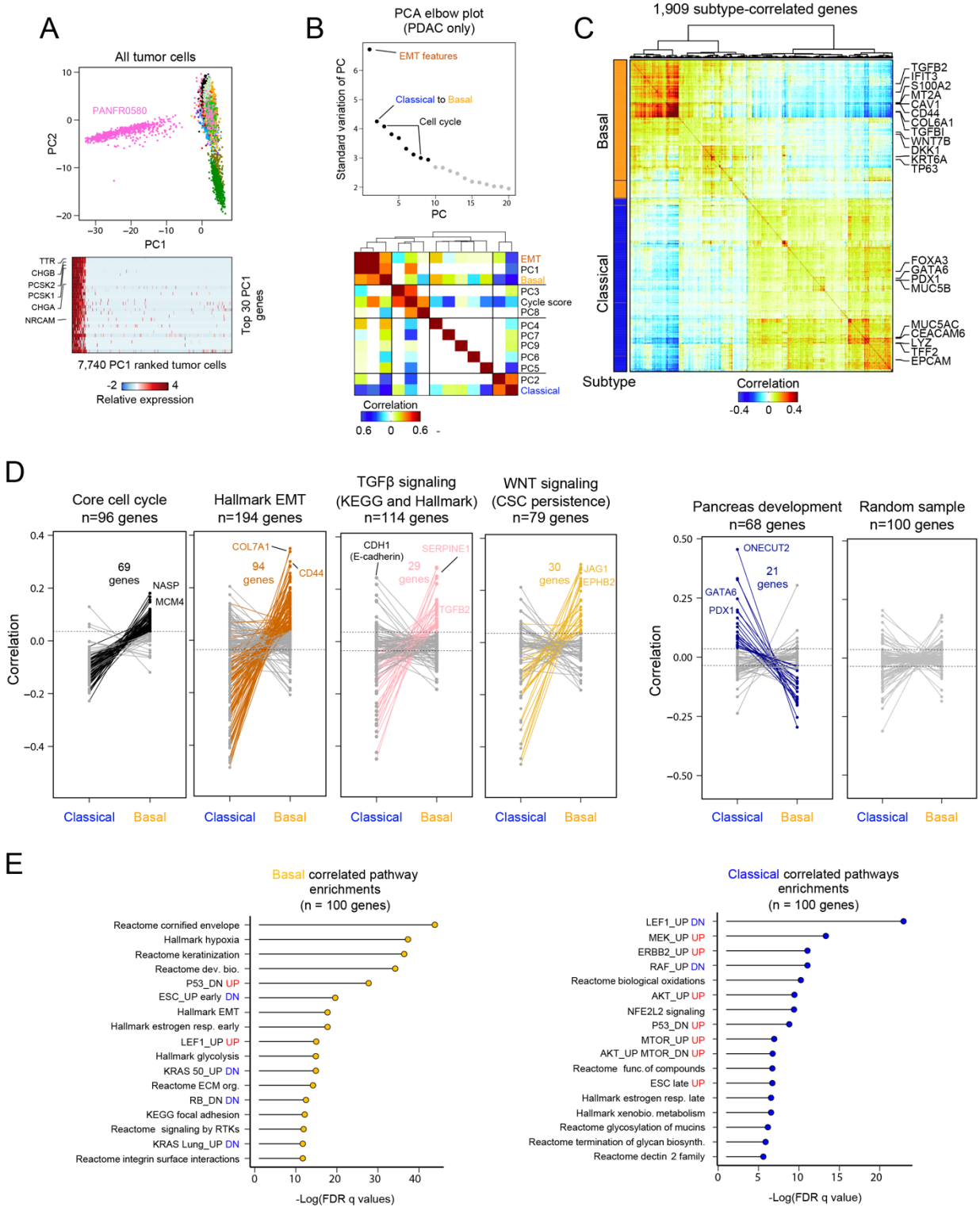

E
Estimated Copy Number
Malignant CNVs
Loss
Gain
Targeted DNA-seq (Bulk)
PANFR0545
PANFR0383
PANFR0598
PANFR0605
PANFR0578
PANFR0557
Sinlge cells
Sub-clone A
Sub-clone B
Chr. 1 10 22
F
Non-malignant cells
Putative malignant cells
T/NK, Fib, Hep, Endo
No malignant cells
CNV Correlation (Avg. top 5% altered cells per tumor)
PANFR0545
PANFR0543
PANFR0604
PANFR0383
PANFR0580
PANFR0526
PANFR0598
PANFR0576
PANFR0489
PANFR0605
PANFR0489R
PANFR0635
PANFR0578
PANFR0562
PANFR0552
PANFR0557
PANFR0583
PANFR0637
CNV correlation
PANFR0575
PANFR0588
PANFR0593
PANFR0592
PANFR0631
CNV Score (mean square of alterations)


G
Non-Malignant
n =15,302 cells
Malignant
n =7,740 cells
Cell type markers
CD14
CD163
CD1D
LILRA4
CLEC4C
XCR1
IL7R
CD3D
CD8A
FOXP3
Tumor keratins
MS4A1
ALB
KDR
FN1
ACTA2
KRT7
KRT19
Cell types
CNV Score
0.002
0.001
0
Cell cycle
Biopsy ID
545
383
580
605
562
598
552
575
576
578
557
588
631
583
635
592
489
526
489R
543
637
593
604
All biopsy cells ranked by SNN cluster (n=23,042)
Cell cycle
Cycling
Cell types
Mac
cDC2
pDC
XCR1+ DC
T/NK
T Reg
B
Plasma
Hep
Endo
Mes
Malignant
Expression
ln(scaled(UMI+1))
0
3.5
7
H
Cell type ID
Malignant cells
T Reg
T/NK
Plasma
Endo
XCR1+DC
cDC2
Mac
pDC
B
Hep
Mes
I
Cell types by patient
n cells:
5,720
5,084
871
759
683
571
475
463
271
259
146
7,740
Fraction of cell type
1.00
0.75
0.50
0.25
0.00
T/NK
Mac
Mes
cDC2
B
Plasma
T Reg
Endo
Hep
pDC
XCR1+ DC
Malignant

Supplemental Figure S2.1. Quality metrics, unsupervised cell type identification, and malignant cell confirmation across the biopsy cohort.

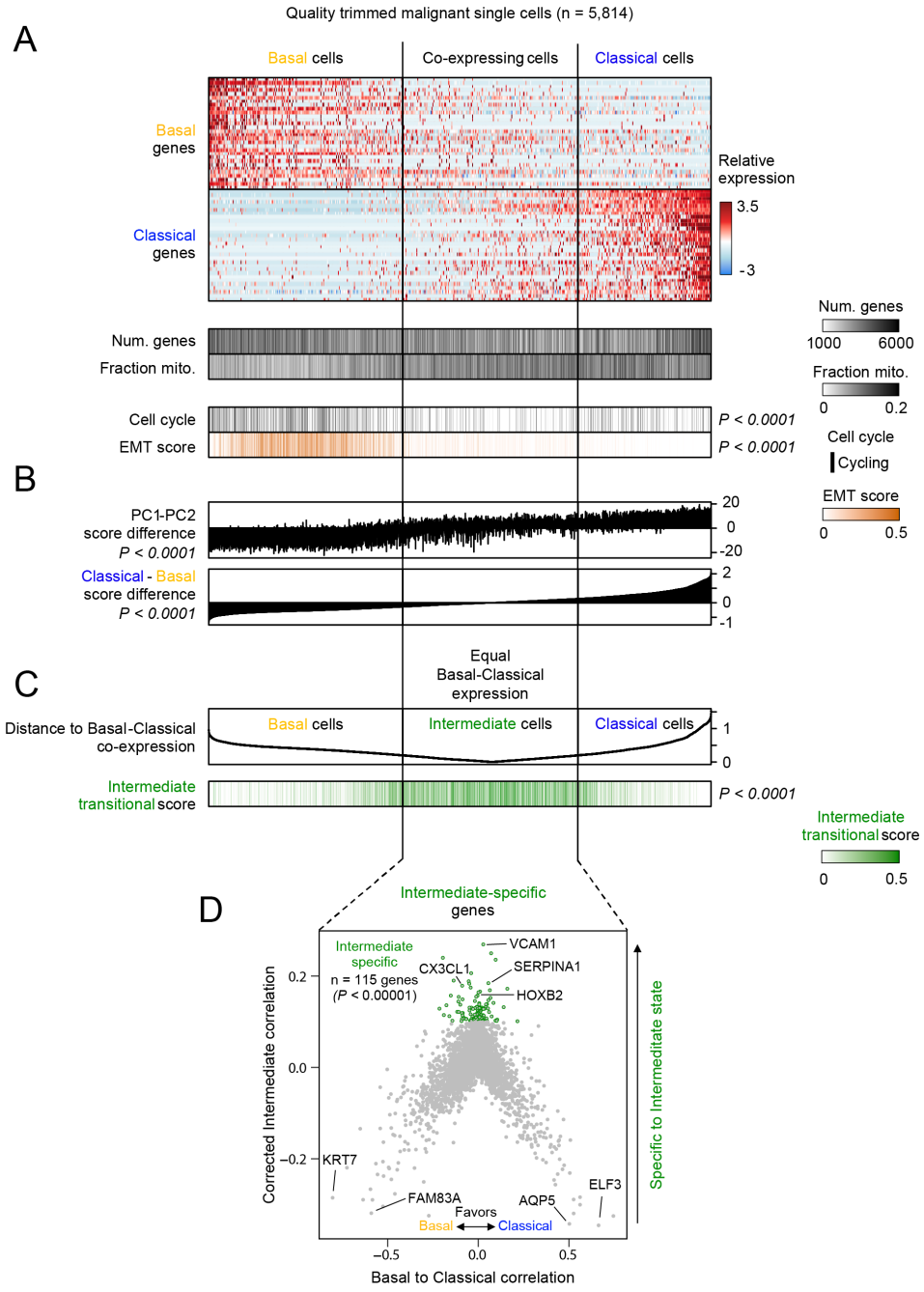
Related to Figure 2.1

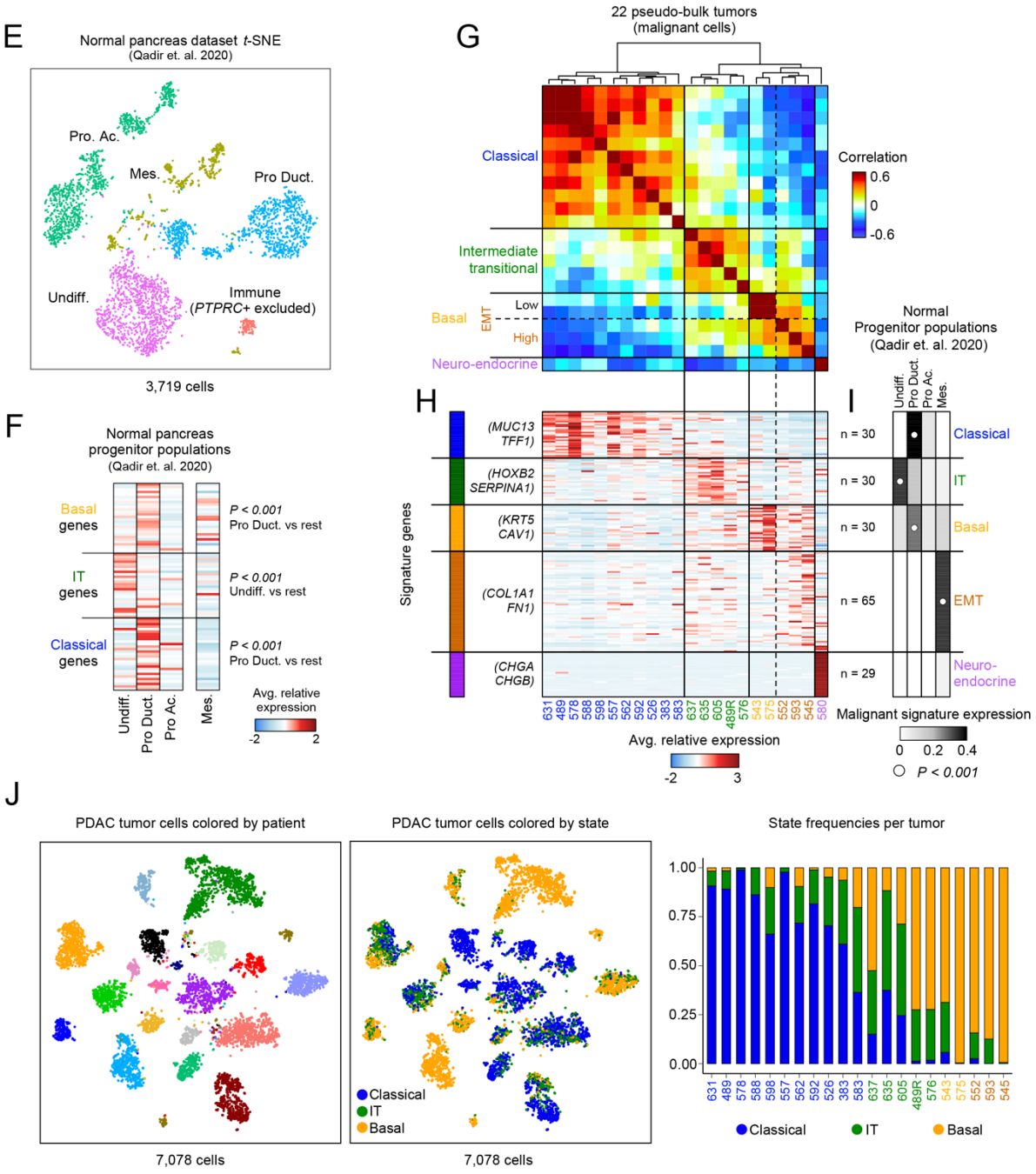
(A) Distribution of unique molecules and genes captured in quality cells per biopsy, median values are indicated for each metric (dotted line) and violin plots are colored by patient (top, Log_{10}(UMIs); bottom, number of genes). **(B)** Distribution of fraction mitochondrial reads across the entire trimmed biopsy dataset (n = 23,042 cells). Red dotted line denotes the median. **(C)** *t*-SNE visualization of the entire single-cell biopsy dataset colored by the SNN clusters identified (inset numbers). **(D)** Distribution of single cells captured per biopsy across the identified SNN clusters. In general, a patient's malignant cells are expected to form unique clusters driven by CNVs. Owing to this feature, the data are split into putative malignant and non-malignant groups of clusters. **(E)** Heatmaps represent select scRNA-seq-derived copy number profiles where expression across the transcriptome is organized by chromosome (columns) for each single putative malignant cell (rows) from a given biopsy. Top bar indicates reference bulk targeted DNA-seq for the same patient and shows strong concordance with the single-cell derived profiles. **(F)** CNV correlation (averaged top 5% of altered cells per biopsy) versus CNV score (mean square of modified expression) for each single putative malignant (colored points) and reference normal cell (empty black circles) within a given biopsy. Only a single sample, PANFR0604, did not contain any malignant cells. **(G)** Overview of cell-typing for all cells in the biopsy dataset. Cells are ordered by SNN cluster and separated by cell types. Top heatmap represent expression levels for a subset of select markers (n=73 genes) used to identify cell types. Color bar indicates cell types and binarized cell cycle phenotypes are labeled (black, cycling; white, not). CNV scores (mean square of alterations per cell) used to parse malignant from non-malignant are shown using T/NK, endothelial, fibroblasts, and hepatocytes as reference; grey boxes denote normal cell types where we did not compute reference CNV scores. Bottom panel shows biopsy of origin for each cell. The data are split by non-malignant (n = 15,302) and malignant (7,740) identity. **(H)** *t*-SNE visualization as in **S2.1C** but colored by cell types identified**,** abbreviations as in **Figure 2.1D**. **(I)** Fraction of each cell type contributed by each biopsy sample (color fill, patient ID; as in **Figure 2.1B**), cell type totals are noted at the top of each bar.


A
All tumor cells
PANFR0580
PC2
PC1
TTR
CHGB
PCSK2
PCSK1
CHGA
NRCAM
Top 30 PC1 genes
7,740 PC1 ranked tumor cells
Relative expression
B
PCA elbow plot (PDAC only)
EMT features
Classical to Basal
Cell cycle
Standard variation of PC
PC
EMT
PC1
Basal
PC3
Cycle score
PC8
PC4
PC7
PC9
PC6
PC5
PC2
Classical
Correlation
C
1,909 subtype-correlated genes
Basal
Classical
Subtype
TGFB2
IFIT3
S100A2
MT2A
CAV1
CD44
COL6A1
TGFBI
WNT7B
DKK1
KRT6A
TP63
FOXA3
GATA6
PDX1
MUC5B
MUC5AC
CEACAM6
LY2
TFF2
EPCAM
Correlation
D
Core cell cycle n=96 genes
Hallmark EMT n=194 genes
TGFβ signaling (KEGG and Hallmark) n=114 genes
WNT signaling (CSC persistence) n=79 genes
Pancreas development n=68 genes
Random sample n=100 genes
Correlation
69 genes
NASP
MCM4
COL7A1
94 genes
CD44
CDH1 (E-cadherin)
SERPINE1
29 genes
TGFB2
30 genes
JAG1
EPHB2
ONECUT2
21 genes
GATA6
PDX1
Classical
Basal
E
Basal correlated pathway enrichments (n = 100 genes)
Reactome cornified envelope
Hallmark hypoxia
Reactome keratinization
Reactome dev. bio.
P53_DN UP
ESC_UP early DN
Hallmark EMT
Hallmark estrogen resp. early
LEF1_UP UP
Hallmark glycolysis
KRAS 50_UP DN
Reactome ECM org.
RB_DN DN
KEGG focal adhesion
Reactome signaling by RTKs
KRAS Lung_UP DN
Reactome integrin surface interactions
-Log(FDR q values)
Classical correlated pathways enrichments (n = 100 genes)
LEF1_UP DN
MEK_UP UP
ERBB2_UP UP
RAF_UP DN
Reactome biological oxidations
AKT_UP UP
NFE2L2 signaling
P53_DN UP
MTOR_UP UP
AKT_UP MTOR_DN UP
Reactome func.of compounds
ESC late UP
Hallmark estrogen resp. late
Hallmark xenobio. metabolism
Reactome glycosylation of mucins
Reactome termination of glycan biosynth.
Reactome dectin 2 family
-Log(FDR q value)

Supplemental Figure S2.2. Identifying and contextualizing basal and classical associated biology.
Related to Figures 2.1 & 2.2

(A) Principal component analysis (PCA) and scatter plot for PC1 and PC2 across all malignant cells (n=7,740) separates PANFR0580's malignant cells (n=662) from the rest of the samples. Cells are colored by patient ID (as in **Figure 2.1B**). Heatmap for genes with the strongest negative loading on PC1 (n=30) denote a neuroendocrine identity (*TTR, CHGB*). This tumor was later classified by histology as a pancreatic neuroendocrine tumor (PanNET). **(B)** Principal component (PC) elbow plot showing the standard deviation for the first 20 components calculated over the verified PDAC malignant cell variable genes (**Methods**). Line is drawn at the putative "elbow" (black versus grey points) as inclusion of additional PCs described overlapping information or quality metrics. Cross-correlational analysis for each single-cell's embeddings across first 9 PCs (black points) and scores for literature curated gene sets describing EMT, classical and basal, and cell cycle phenotypes. PC1 positively correlates with EMT, basal, and to a degree, cell cycling. Cells with positive embeddings on PC2 are correlated with classical phenotypes and anti-correlated with basal and EMT phenotypes, suggesting these phenotypes are anti-correlated across a continuum of expression. PC3 and PC8 describe cells with high cell cycle scores. The other PCs do not associate significantly with these phenotypes. **(C)** Pairwise correlation of genes significantly associated with basal (PC1/negative PC2) or classical (PC2) expression states. Left bar indicates the subtype association of each gene (orange, basal; blue, classical). **(D)** Tied dot plots depicting the correlation coefficient for each gene (points) to either basal or classical phenotypes from select literature-derived gene sets, indicated at the top of each plot, which summarize aspects of subtype associated biology. Dotted lines represent significance threshold (3 SD above the mean of shuffled data), points and lines are colored if that gene passes the threshold and select genes are indicated. **(E)** GSEA pathway enrichments for top 100 genes correlated to either basal or classical expression scores.

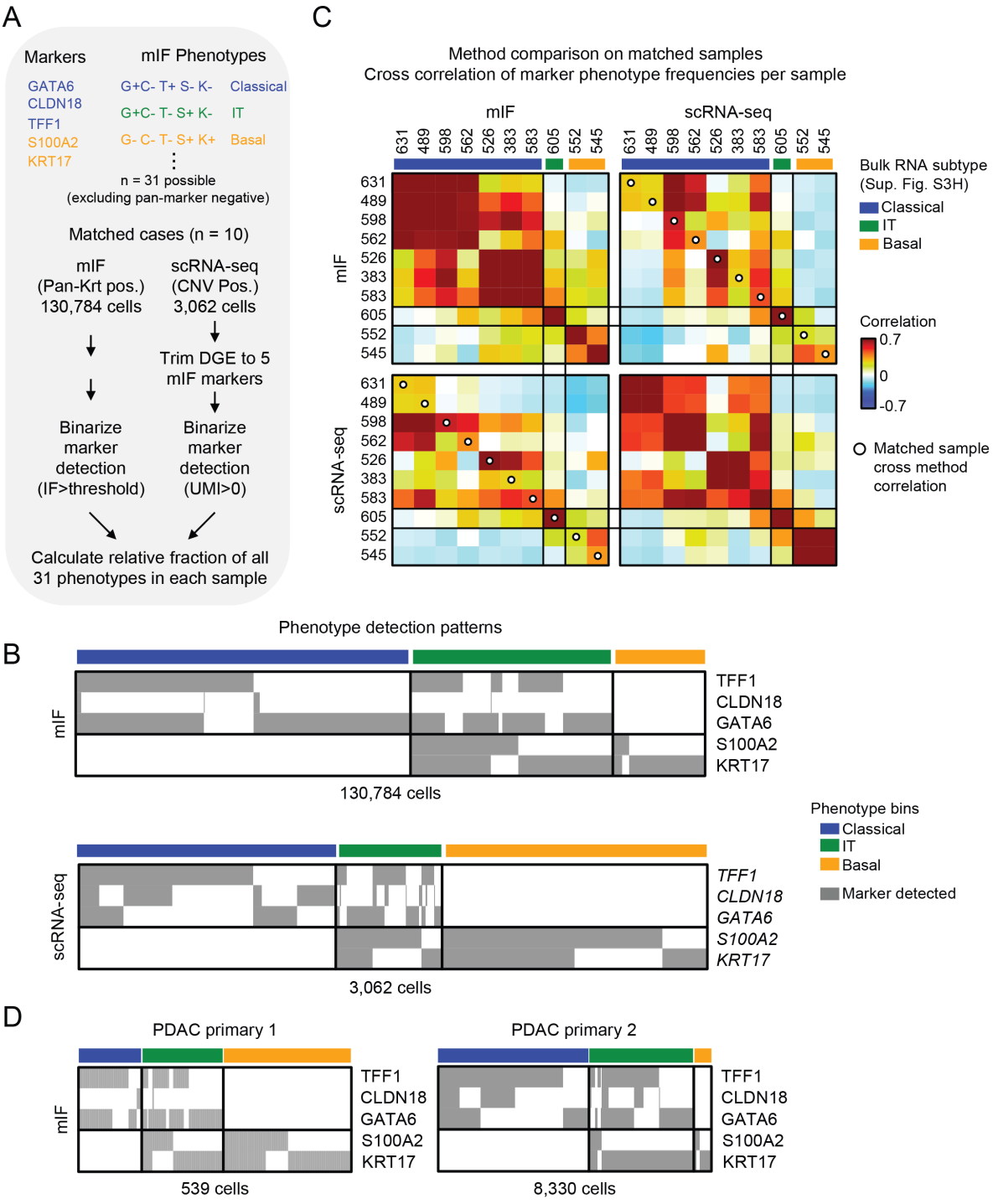

A
Quality trimmed malignant single cells (n = 5,814)
Basal cells
Co-expressing cells
Classical cells
Basal genes
Classical genes
Relative expression
3.5
-3
Num. genes
Fraction mito.
Cell cycle
EMT score
P < 0.0001
P < 0.0001
Num. genes
1000 6000
Fraction mito.
0 0.2
Cell cycle
Cycling
EMT score
0 0.5
B
PC1-PC2 score difference
P < 0.0001
20
0
-20
Classical - Basal score difference
P < 0.0001
2
0
-1
C
Equal Basal-Classical expression
Distance to Basal-Classical co-expression
Basal cells
Intermediate cells
Classical cells
1
0
Intermediate transitional score
P < 0.0001
Intermediate transitional score
0 0.5
D
Intermediate-specific genes
Intermediate specific
n = 115 genes
(P < 0.00001)
VCAM1
CX3CL1
SERPINA1
HOXB2
KRT7
FAM83A
Favors
Basal ↔ Classical
AQP5
ELF3
Corrected Intermediate correlation
0.2
0.0
−0.2
−0.5
0.0
0.5
Basal to Classical correlation
Specific to Intermediate state


E
Normal pancreas dataset t-SNE
(Qadir et. al. 2020)
Pro. Ac.
Mes.
Pro Duct.
Undiff.
Immune
(PTPRC+ excluded)
3,719 cells
F
Normal pancreas
progenitor populations
(Qadir et. al. 2020)
Basal genes
IT genes
Classical genes
P < 0.001
Pro Duct. vs rest
P < 0.001
Undiff. vs rest
P < 0.001
Pro Duct. vs rest
Undiff.
Pro Duct.
Pro Ac.
Mes.
Avg. relative expression
-2
2
G
22 pseudo-bulk tumors
(malignant cells)
Classical
Intermediate transitional
Basal
EMT
Low
High
Neuro-endocrine
Correlation
0.6
0
-0.6
Normal
Progenitor populations
(Qadir et. al. 2020)
H
Signature genes
(MUC13 TFF1)
(HOXB2 SERPINA1)
(KRT5 CAV1)
(COL1A1 FN1)
(CHGA CHGB)
631 489 578 588 598 557 562 592 526 383 583 637 635 605 489R 576 543 575 552 593 545 580
Avg. relative expression
-2
3
I
Undiff.
Pro Duct.
Pro Ac.
Mes.
n = 30
n = 30
n = 30
n = 65
n = 29
Classical
IT
Basal
EMT
Neuro-endocrine
Malignant signature expression
0 0.2 0.4
P < 0.001
J
PDAC tumor cells colored by patient
7,078 cells
PDAC tumor cells colored by state
Classical
IT
Basal
7,078 cells
State frequencies per tumor
1.00
0.75
0.50
0.25
0.00
631 489 578 588 598 557 562 592 526 383 583 637 635 605 489R 576 543 575 552 593 545
Classical
IT
Basal

Supplemental Figure S2.3. Cells with intermediate co-expressing phenotypes express a distinct gene program.

Related to Figure 2.2

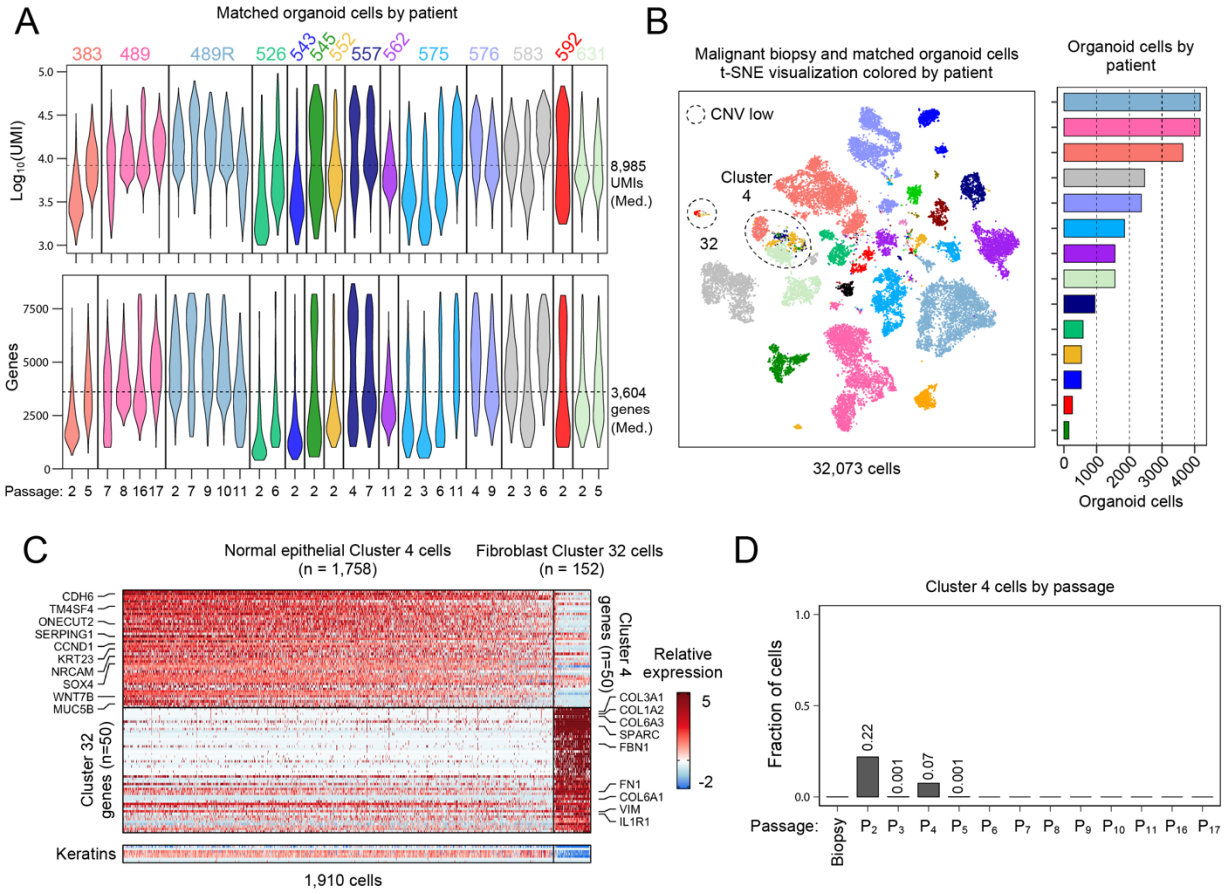
(A) Expression of basal and classical gene programs, with cells ordered by their basal-classical score difference. Quality metrics, EMT scores and the binarized cell cycle program are shown for each single cell below the heatmap. **(B)** PC1 and 2 difference (top) and Classical – Basal score difference (bottom) are shown. Cells with equal basal and classical expression are associated with intermediate PC scores and cells are ordered as in **S2.3A**. **(C)** Euclidian distance for each cell to co-expression (y = x) of basal (x) and classical (y) expression scores. Bottom track indicates the score derived from the genes specific to the intermediate state shown in **S2.3D** and explained in **Methods**. **(D)** Gene correlation to either basal or classical score (x axis) or the corrected intermediate correlation (Euclidean distance in **S2.3C**, **Methods**). Green highlighted genes have corrected intermediate correlation >0.1 ($P<0.00001$ above shuffled). *P*-value for binarized cycling group differences in **S2.3A** was calculated using Fisher's Exact test. *P*-values for EMT score in **S2.3A** and group differences in **S2.3B** and **S2.3C** were calculated by Kruskal-Wallis test with multiple hypothesis correction. **(E)** *t*-SNE visualization after dimensionality reduction and re-clustering for the normal progenitor populations identified in Qadir et al., 2020. Cell types are collapsed to those favoring Acinar (Pro Ac.), Ductal (Pro Duct.), or Undifferentiated (Undiff.) subsets. Mesenchymal cells (Mes.) are included as a non-epithelial reference and the small subset of immune cells was excluded from the comparisons. **(F)** Averaged expression of all three malignant programs in normal pancreatic progenitor niche subsets and mesenchymal cells defined in Qadir et al., 2020. *P*-values for each set of genes are computed by Kruskal-Wallis test with multiple hypothesis correction. **(G)** Pairwise correlation for biopsies with malignant cells (n = 22). Data are correlation coefficients for the average expression of all signature genes in the malignant cells from a given biopsy. EMT genes are from Groger et al., 2012. Clade identities are at left with the one PanNET tumor (PANFR0580) included for comparison and PANFR0604 not included due to lack of malignant cells captured. **(H)** Average expression for the 184 genes used for clustering in **S2.3G**. Clade identity colors match text color in **S2.3G** and individual samples (columns) are ordered as in **S2.3G** and sample ID numbers are provided below. **(I)** Scores for the expression of genes in **S2.3H** (grey scale heat) across the 4 main cell types found in the pancreatic progenitor niche (Qadir et al., 2020). White dot indicates the normal subset with the highest average expression for each malignant program (Kruskal-Wallis test), none of the normal subsets significantly express the Neuroendocrine gene signature. **(J)** *t*-SNE visualization for malignant single cells in the biopsy cohort demonstrates intratumoral transcriptional heterogeneity at the single-cell level. Cells are colored by patient (left) or by transcriptional subtype (right).

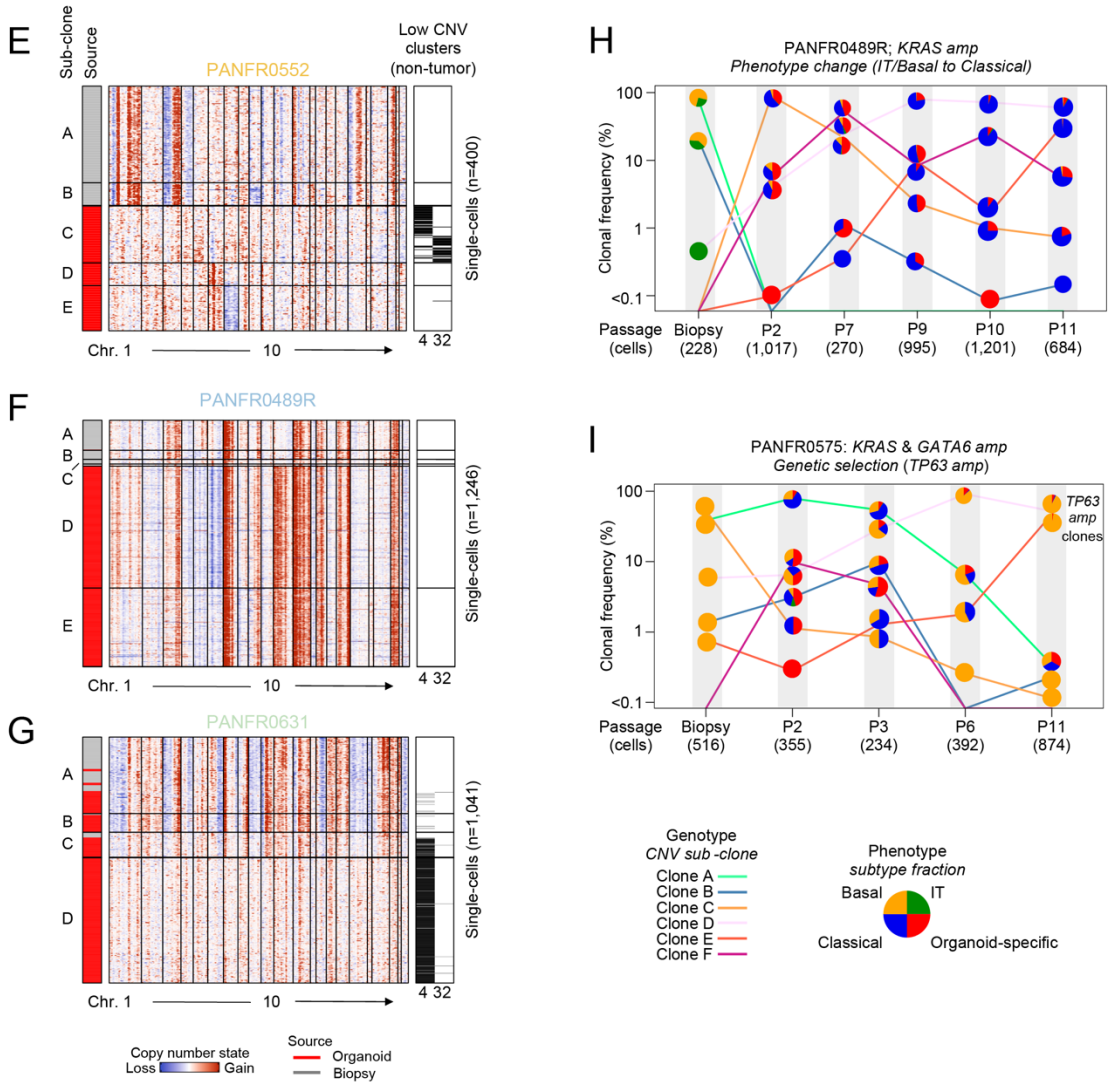

A
Markers
mIF Phenotypes
GATA6
CLDN18
TFF1
S100A2
KRT17
G+C- T+ S- K-
Classical
G+C- T- S+ K-
IT
G- C- T- S+ K+
Basal
n = 31 possible
(excluding pan-marker negative)
Matched cases (n = 10)
mIF
(Pan-Krt pos.)
130,784 cells
scRNA-seq
(CNV Pos.)
3,062 cells
Trim DGE to 5
mIF markers
Binarize
marker
detection
(IF>threshold)
Binarize
marker
detection
(UMI>0)
Calculate relative fraction of all
31 phenotypes in each sample
C
Method comparison on matched samples
Cross correlation of marker phenotype frequencies per sample
mIF
scRNA-seq
631
489
598
562
526
383
583
605
552
545
Bulk RNA subtype
(Sup. Fig. S3H)
Classical
IT
Basal
Correlation
0.7
0
-0.7
Matched sample
cross method
correlation
B
Phenotype detection patterns
mIF
TFF1
CLDN18
GATA6
S100A2
KRT17
130,784 cells
scRNA-seq
TFF1
CLDN18
GATA6
S100A2
KRT17
3,062 cells
Phenotype bins
Classical
IT
Basal
Marker detected
D
PDAC primary 1
PDAC primary 2
mIF
TFF1
CLDN18
GATA6
S100A2
KRT17
539 cells
8,330 cells

Supplemental Figure S2.4. Multiplex immunofluorescence is concordant with scRNA-seq and demonstrates intratumoral heterogeneity with the presence of IT cells.

Related to Figure 2.2

(A) Schematic for comparison of the matched datasets by combinatorial marker phenotypes. **(B)** Marker detection in each single cell from the 10 samples in the mIF (top, 130,784 cells) and matched scRNA-seq datasets (bottom, 3,062 cells). Cells are sorted by their combinatorial phenotype outlined in **S2.4A**. **(C)** Comparison within and between modalities on matched samples. Samples are sorted by the dendrogram in **Supplemental Figure S2.3G** and labeled with their pseudo-bulk RNA subtype identity. Correlation is performed over the fractional representation of each mIF phenotype (**S2.4A**) in each biopsy. Despite measuring different molecules (protein vs mRNA), the two approaches were highly concordant within RNA subtypes and on a case-by-case basis (white dots). (**D**) mIF marker detection in each single cell from two primary PDAC samples shown in **Figure 2.2H**. Cells are sorted by their combinatorial phenotype outlined in **S2.4A**.


A
Matched organoid cells by patient
383 489 489R 526 543 545 552 557 562 575 576 583 592 631
Log10(UMI)
8,985 UMIs (Med.)
Genes
3,604 genes (Med.)
Passage: 2 5 7 8 16 17 2 7 9 10 11 2 6 2 2 2 4 7 11 2 3 6 11 4 9 2 3 6 2 2 5
B
Malignant biopsy and matched organoid cells t-SNE visualization colored by patient
CNV low
Cluster 4
32
32,073 cells
Organoid cells by patient
Organoid cells
C
Normal epithelial Cluster 4 cells (n = 1,758)
Fibroblast Cluster 32 cells (n = 152)
CDH6
TM4SF4
ONECUT2
SERPING1
CCND1
KRT23
NRCAM
SOX4
WNT7B
MUC5B
Cluster 32 genes (n=50)
Cluster 4 genes (n=50)
COL3A1
COL1A2
COL6A3
SPARC
FBN1
FN1
COL6A1
VIM
IL1R1
Relative expression
Keratins
1,910 cells
D
Cluster 4 cells by passage
Fraction of cells
0.22
0.001
0.07
0.001
Passage: Biopsy P2 P3 P4 P5 P6 P7 P8 P9 P10 P11 P16 P17

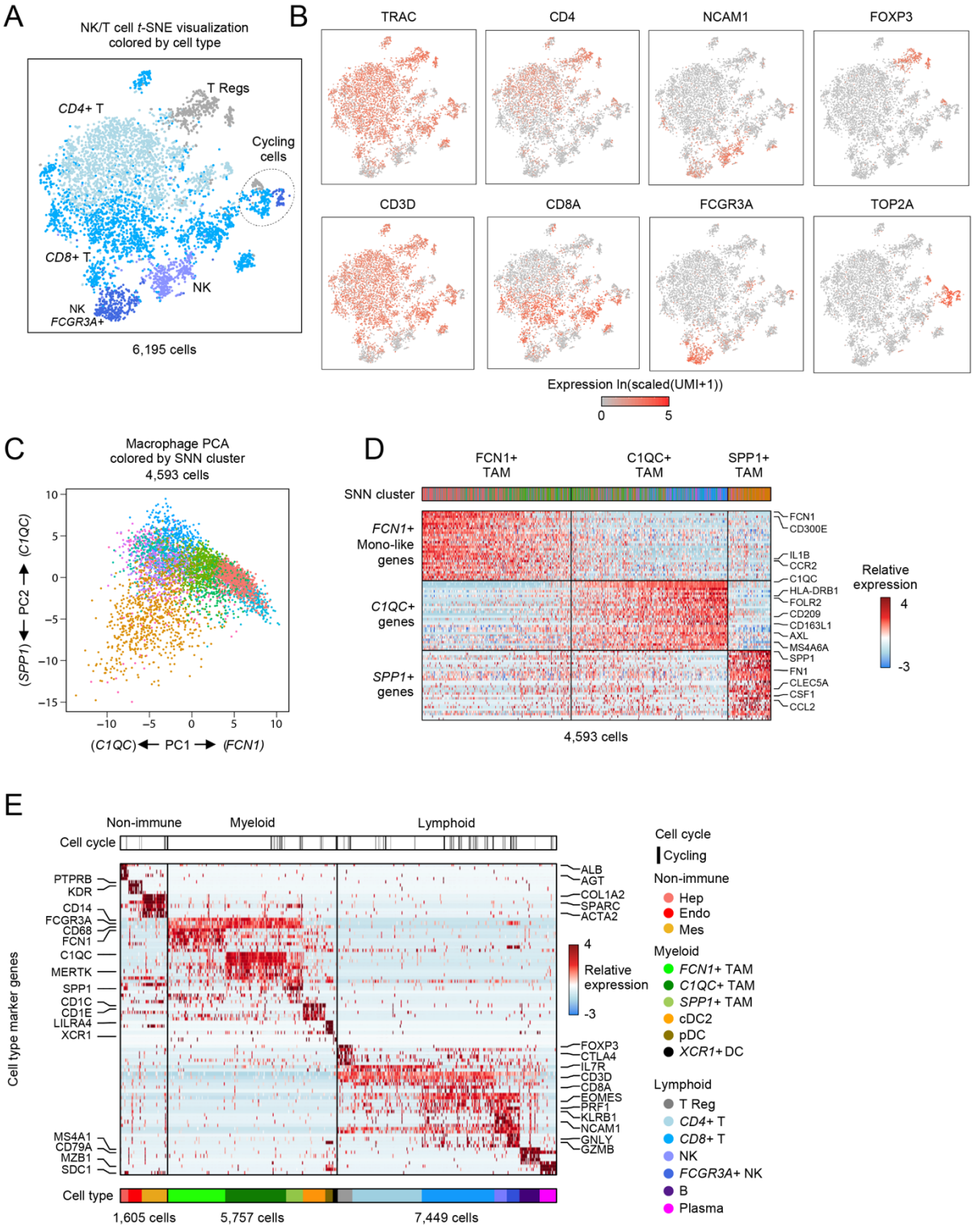


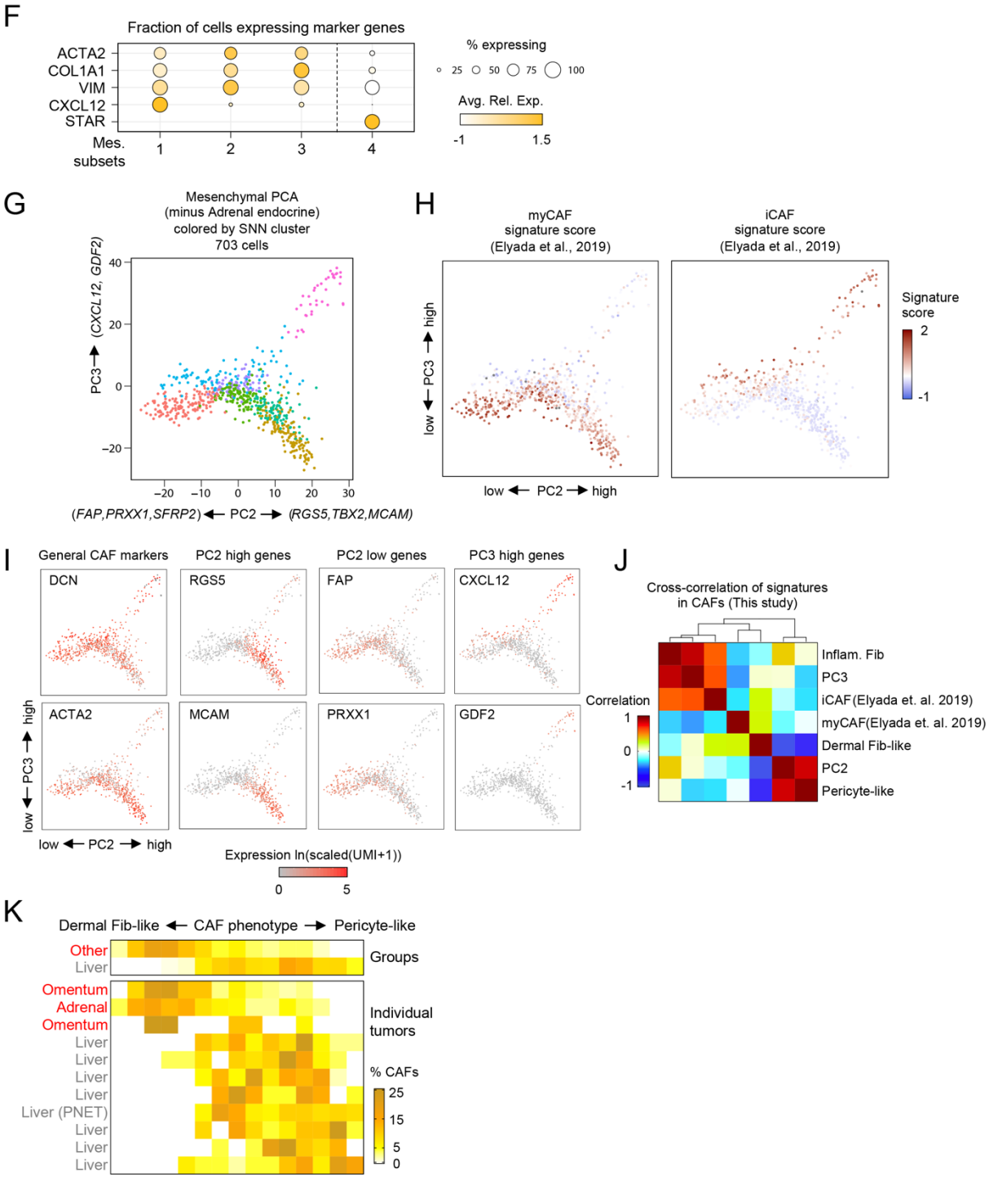
Supplemental Figure S2.5. Quality metrics, cell type identification, and serial sampling across the patient-matched organoid cohort.

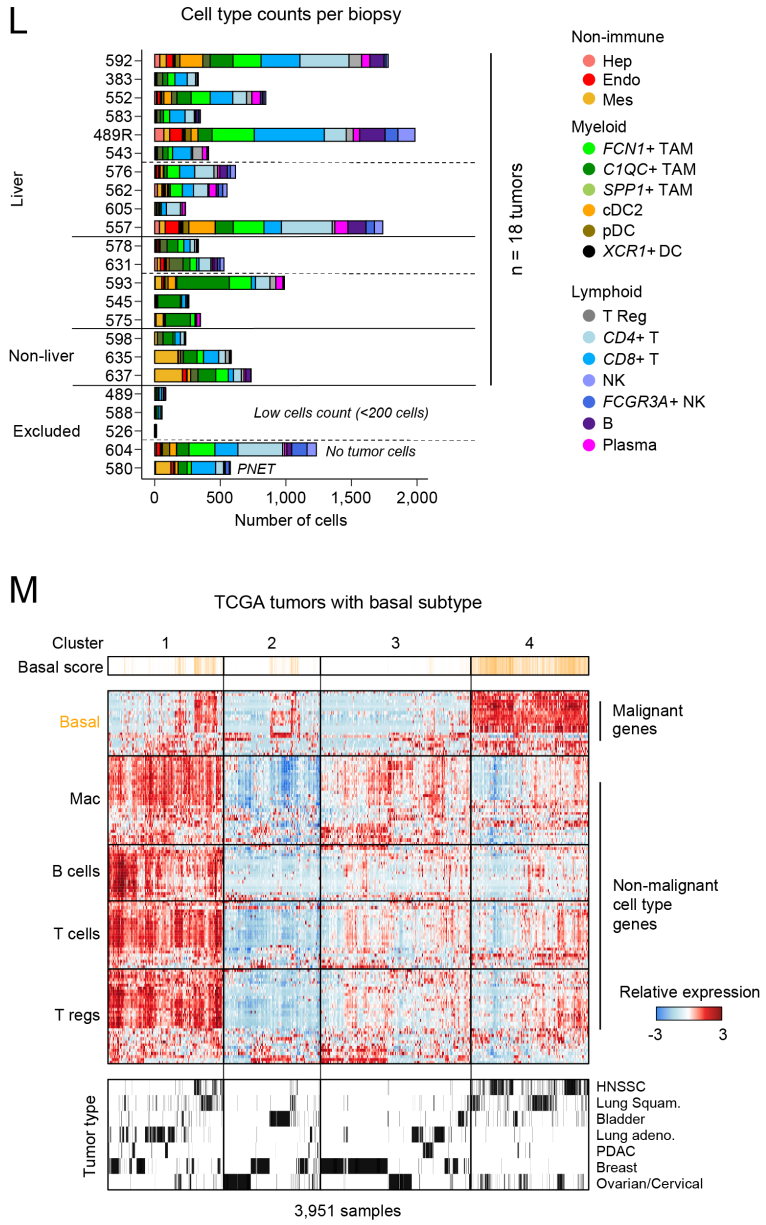
Related to Figure 2.4

(A) Distribution of unique molecules and genes captured in quality cells per organoid sample, median values are indicated for each metric (dotted line) and violin plots are colored by patient ID (top, Log_{10}(UMIs); bottom, number of genes). **(B)** *t*-SNE visualization of all biopsy and matched organoid cells from iterative passages, colored by patient ID. Dotted circles indicate the only two SNN clusters (4 and 32) with appreciably admixed clusters and low CNV scores, the rest were patient-specific. Bar chart shows number of organoid cells recovered per matched sample (right). **(C)** Relative expression for genes defining cluster 4 (top) and cluster 32 (bottom; 1 versus rest DE with the cells in **S2.5B**). Cluster 4 had an ambiguous epithelial identity while cluster 32 cells were defined by canonical fibroblast genes and low to absent detection of CNVs. **(D)** Fraction of cluster 4 cells at each passage. These cells did not survive iterative

passaging suggesting that they were either untransformed or unfit in organoid culture. **(E-G**) Heatmaps show inferred CNV copy number status for every cell in each of three biopsy/early passage organoid pairs. Cells are ordered by hierarchical clustering of their CNV profiles and letters on the far left indicate subclones that have significant statistical evidence for tree-splitting (**Methods**). Each cell's origin (biopsy tissue, grey; early passage organoid, red) is also noted ("Source" column). Right metadata bars indicate if that cell came from an admixed SNN cluster (4 or 32 in **S2.5B**). (**H**, **I**) Matched phenotype and genotype evolution at each passage in PANFR0489R (**S2.5H**) and PANFR0575 (**S2.5I**). Frequencies of individual CNV clones at each time point (**Methods**, y axis) are tied by colored lines. Fill represents the transcriptional phenotype fraction for each CNV clone. In sample PANFR0575 (**S2.5I**), clones D and E had inferred *TP63* amplifications which expanded over time.


A
NK/T cell t-SNE visualization
colored by cell type
T Regs
CD4+ T
Cycling
cells
CD8+ T
NK
NK
FCGR3A+
6,195 cells
B
TRAC
CD4
NCAM1
FOXP3
CD3D
CD8A
FCGR3A
TOP2A
Expression ln(scaled(UMI+1))
0
5
C
Macrophage PCA
colored by SNN cluster
4,593 cells
(SPP1) ← PC2 → (C1QC)
(C1QC) ← PC1 → (FCN1)
D
FCN1+
TAM
C1QC+
TAM
SPP1+
TAM
SNN cluster
FCN1+
Mono-like
genes
C1QC+
genes
SPP1+
genes
FCN1
CD300E
IL1B
CCR2
C1QC
HLA-DRB1
FOLR2
CD209
CD163L1
AXL
MS4A6A
SPP1
FN1
CLEC5A
CSF1
CCL2
Relative
expression
4
-3
4,593 cells
E
Non-immune
Myeloid
Lymphoid
Cell cycle
Cell type marker genes
PTPRB
KDR
CD14
FCGR3A
CD68
FCN1
C1QC
MERTK
SPP1
CD1C
CD1E
LILRA4
XCR1
MS4A1
CD79A
MZB1
SDC1
ALB
AGT
COL1A2
SPARC
ACTA2
FOXP3
CTLA4
IL7R
CD3D
CD8A
EOMES
PRF1
KLRB1
NCAM1
GNLY
GZMB
4
Relative
expression
-3
Cell type
1,605 cells
5,757 cells
7,449 cells
Cell cycle
Cycling
Non-immune
Hep
Endo
Mes
Myeloid
FCN1+ TAM
C1QC+ TAM
SPP1+ TAM
cDC2
pDC
XCR1+ DC
Lymphoid
T Reg
CD4+ T
CD8+ T
NK
FCGR3A+ NK
B
Plasma


F
Fraction of cells expressing marker genes
ACTA2
COL1A1
VIM
CXCL12
STAR
Mes. subsets
1
2
3
4
% expressing
25
50
75
100
Avg. Rel. Exp.
-1
1.5
G
Mesenchymal PCA
(minus Adrenal endocrine)
colored by SNN cluster
703 cells
PC3 (CXCL12, GDF2)
(FAP,PRXX1,SFRP2) PC2 (RGS5,TBX2,MCAM)
H
myCAF
signature score
(Elyada et al., 2019)
iCAF
signature score
(Elyada et al., 2019)
low PC3 high
low PC2 high
Signature score
2
-1
I
General CAF markers
PC2 high genes
PC2 low genes
PC3 high genes
DCN
RGS5
FAP
CXCL12
ACTA2
MCAM
PRXX1
GDF2
low PC3 high
low PC2 high
Expression ln(scaled(UMI+1))
0
5
J
Cross-correlation of signatures
in CAFs (This study)
Correlation
1
0
-1
Inflam. Fib
PC3
iCAF(Elyada et. al. 2019)
myCAF(Elyada et. al. 2019)
Dermal Fib-like
PC2
Pericyte-like
K
Dermal Fib-like CAF phenotype Pericyte-like
Other
Liver
Groups
Omentum
Adrenal
Omentum
Liver
Liver
Liver
Liver
Liver (PNET)
Liver
Liver
Liver
Individual tumors
% CAFs
25
15
5
0

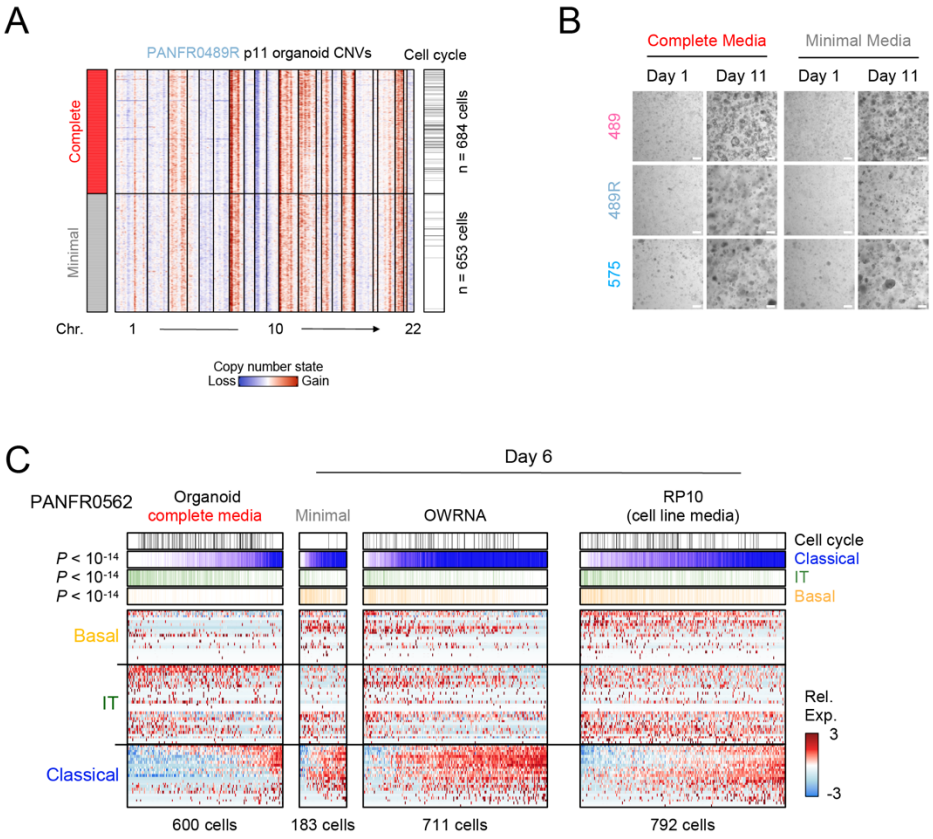


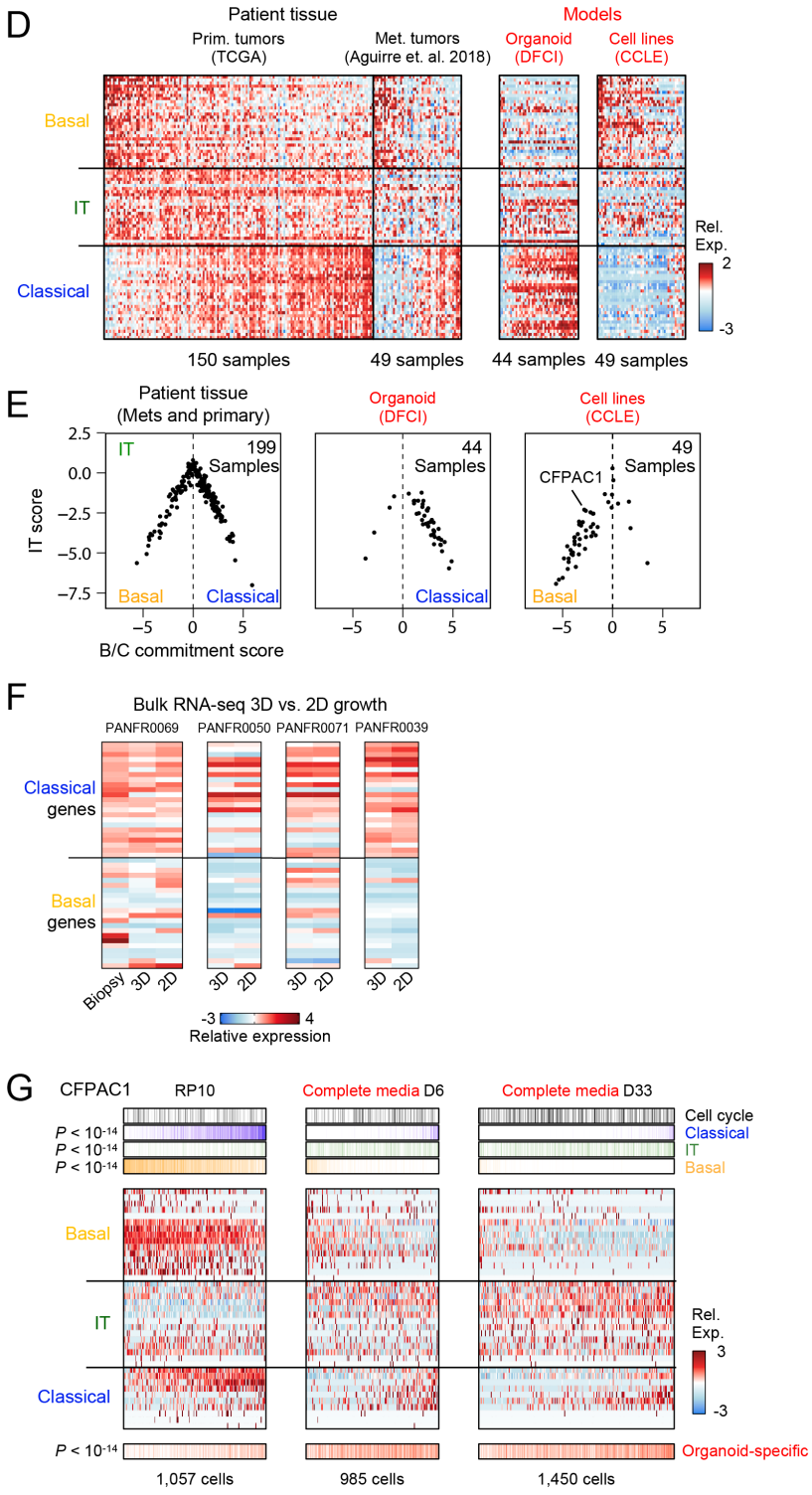
Supplemental Figure S2.6. Identification of T/NK, macrophage, and fibroblast heterogeneity in the metastatic microenvironment.

Related to Figures 2.5 & 2.6

(A) *t*-SNE visualization of sub-clustering (SNN) performed on T/NK cells in the metastatic cohort. Cells are colored by their type identity based on shared SNN cluster membership (Methods). **(B)** Select cell type marker expression overlaid on the *t*-SNE visualization from **S2.6A. (C)** PCA identifies 3 major subsets of TAMs in the metastatic niche. PC1 largely separates *FCN1*+ monocyte-like TAMs from more committed macrophage phenotypes. PC2 separates *SPP1*+ from *C1QC*+ macrophage phenotypes. **(D)** Heatmap visualization of the gene expression programs specific to each TAM subset identified by the PCA in **S2.6C**

(**Methods**). Top metadata indicate SNN cluster as in **S2.6C**. **(E)** Heatmap shows the relative expression for select cell type markers. Top bar indicates the binarized cell cycle program (black, cycling) and the bottom color bar corresponds to the cell type colors noted in **Figure 2.6A. (F)** Dot plots for average expression of the indicated CAF and adrenal endocrine marker genes in each of the cell subsets (1-4) identified in **Figure 2.5C**. Size of the dot indicates fraction of cells expressing a given gene. **(G)** PCA over fibroblasts in the cohort (excluding Adrenal endocrine cells; subset 4, **Figure 2.5C**). Scatter plot of PC2 vs PC3 defines 3 states for CAFs in our cohort (**Methods**). **(H)** Same visualization in **S2.6B**, but cells are colored by previously identified myCaf or iCaf signature scores. myCaf is evenly distributed across PC2 and iCaf associates with higher PC3 scores. **(I)** Expression for select markers overlaid on the PCA from **S2.6B**. **(J)** Cross-correlation of fibroblast signatures in single-cells. New dermal- vs. pericyte-like signatures provide non-overlapping information. PC3 inflammatory phenotypes are similar to the previously reported iCaf phenotype (Elyada et al., 2019) and our PC3-derived inflammatory fibroblast signature. **(K)** Distribution across the CAF continuum comparing site differences as groups (top) or individual tumors (bottom). Heat indicates the fraction of CAFs in that score bin. **(L)** Bar plot shows the number of non-malignant cells in each biopsy, color fill indicates the number of each cell type captured in that sample. Five biopsies were excluded from the analysis in **Figure 2.6A-C** because they either had low cell capture or were from a tumor with indeterminant malignant transcriptional subtype. Relevant samples are organized as in **Figure 2.6A**. **(M)** Cross TCGA analysis for basal and immune cell type markers in epithelial tumors with known basal subtypes (Cancer Genome Atlas Research et al., 2013). Tumors with strong basal gene expression do not associate with strong immune infiltrates. Clusters were determined by dendrogram splitting and disease type for each sample is indicated below.


A
PANFR0489R p11 organoid CNVs
Cell cycle
Complete
Minimal
n = 684 cells
n = 653 cells
Chr.
1
10
22
Copy number state
Loss
Gain
B
Complete Media
Minimal Media
Day 1
Day 11
Day 1
Day 11
489
489R
575
C
Day 6
PANFR0562
Organoid
complete media
Minimal
OWRNA
RP10
(cell line media)
Cell cycle
Classical
IT
Basal
P < 10-14
P < 10-14
P < 10-14
Basal
IT
Classical
Rel.
Exp.
3
-3
600 cells
183 cells
711 cells
792 cells



Supplemental Figure S2.7. Alterations to organoid media, but not matrix dimensionality, shift transcriptional phenotype.

Related to Figure 2.7

(A) Inferred CNVs for each cell from the PANFR489R samples cultured in either Minimal (grey) or Complete (red) organoid media conditions in **Figure 2.7B**. **(B)** Brightfield images were obtained for organoids grown in standard organoid media ("Complete") or in media without any growth factors ("Minimal") at days 1 and 11 after seeding. **(C)** Single organoid cells from model PANFR0562 (columns) cultured for 6 days in Complete medium, Minimal medium, OWRNA medium, or in RP10 ("cell line" medium, RPMI-1640 with 10% fetal bovine serum) and scored for basal, IT, and classical hierarchy phenotypes (rows). *P*-values for group differences were calculated by ANOVA followed by Tukey's HSD. **(D)** Relative expression for 90 genes representing PDAC state programs across bulk RNA-seq samples from primary resections (TCGA) and metastatic biopsies (Panc-Seq), as well as organoid and cell line (CCLE) models (Aguirre et al., 2018; Barretina et al., 2012; Cancer Genome Atlas Research Network, 2017; Ghandi et al., 2019). **(E)** PDAC malignant state diagrams for average Basal-classical commitment score (x axis) and IT score (y axis) for bulk RNA-seq samples in **S2.7D**. **(F)** Four established models were adapted to 2-dimensional culture in complete organoid media and measured via bulk RNA-seq. Rows indicate expression levels of basal-classical commitment score genes. **(G)** Single cells from the established PDAC cell line CFPAC1 (columns) sampled in RP10 (standard "cell line" medium, RPMI-1640 with 10% fetal bovine serum) or at 2 timepoints in Complete organoid medium and scored for basal, IT, and classical phenotypes (rows). Bottom row indicates single cell organoid-specific gene expression (as described in **Figure 2.4B**) across all three conditions. *P*-values for group differences were calculated by ANOVA followed by Tukey's HSD.