Multiplex single-cell RNA sequencing for chemical genomics and spatial transcriptomics

Sanjay R Srivatsan

A dissertation
submitted in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy

University of Washington

2021

Reading Committee:

Cole Trapnell, Chair

Jay Shendure

Stephen Tapscott

Program Authorized to Offer Degree:
Department of Genome Sciences

University of Washington

Abstract

Multiplex single-cell RNA sequencing for chemical genomics and spatial transcriptomics

Sanjay R Srivatsan

Chair of the Supervisory Committee:
Professor Cole Trapnell
Genome Sciences

Each of us begins life as a single fertilized cell. Following a seemingly predetermined set of cell divisions, the single cell morphs into a rough mass, then a hollowed tube, and finally becomes a recognizable neonatal form. How the information contained within a single cell simultaneously specifies an organism's anatomy, the construction of its organs, and the ability to cogitate on this very question, remains one of biology's open questions. Although centuries of careful experiments devoted to characterizing development have revealed many important genes and mechanisms, the results of these experiments span different model organisms, developmental stages, cell populations and measurement modalities. Integrating this knowledge base into coherent representation requires a cellular scaffold that charts an organism's development over the axes of time and space. Preliminary unified representations of developing organisms (e.g. C. Elegans, Zebrafish and Mouse) have been created by large-scale single cell RNA sequencing (scRNA-seq) efforts. These efforts have characterized the set of intermediates through which differentiating cells transit and have profiled the large number of cell types present in a developing organism. Although scRNA-seq data have proven powerful in cataloging cellular states, they lack crucial context: i) the experimental context afforded by the comparison of multiple conditions (e.g. wild-type vs. perturbation) and ii) a cell's spatial context, a crucial factor driving its behavior. To address these knowledge gaps, over the course of my PhD I have developed two scRNA-seq technologies: 1) sci-Plex, a generalizable strategy to label cell populations and 2) sci-Space, a methodology to record a

cell's spatial position in conjunction with its single cell transcriptome.

(1) First I developed the sci-Plex protocol, an inexpensive and efficient method to label single cells through the chemical fixation of unmodified single stranded oligos to nuclei prior to scRNA-seq library preparation. To demonstrate proof-of-concept of the sci-Plex protocol, I performed a high-throughput, high-content drug screen at single cell resolution in 3 cancer cell lines; effectively conducting 4,500 independent scRNA-seq experiments at once. The resulting dataset enabled characterization of a drug's potency, class, mechanism of action, and the heterogeneity of cellular responses induced upon drug treatment. For example, our scRNA-seq data showed that histone deacetylase inhibitors likely lead to cell death by trapping valuable acetyl molecules on chromatin.

(2) Next, I extended the application of the sci-Plex protocol and developed the sci-Space method to capture spatial information from sectioned tissue. The fast and scalable sci-Space method uses patterned oligonucleotide barcodes in a regular array such that each spot contains a unique set of sequences. Then, to mark each nucleus' coordinates on the grid, the barcodes are stamped onto a tissue section prior to disaggregation and library preparation. To showcase the power of sci-Space, I collected a dataset comprising over 120,000 cells originating from 14 sections of a single E14 mouse embryo. The resulting data uncovers the genes that drive the developing organism's body plan and reveals a widespread migration signature within neurons that form the developing brain. These data also provide a quantitative assessment of how cell state relates to spatial position within the developing embryo. Specifically, our estimates indicate that 25% of the variance in gene expression observed is attributable to spatial position. It is my hope that this technology will power the generation of a unified scaffold of development akin to the reference genome. I believe that such a unified representation will be instrumental in amassing data, accelerating discovery and facilitating translation through the training of machine learning models of cellular state.

TABLE OF CONTENTS

LIST OF FIGURES

ACKNOWLEDGMENTS

I am deeply grateful to the many people it took to make me the person and scientist I am today. I've trained at many institutions prior to coming to the University of Washington. I was first exposed to science at the Morehouse School of Medicine where I first learned how to do molecular biology experiments as a high schooler. At UC Berkeley in the lab of Steve Martin, Venice Chiueh gave me another chance to become work as a molecular biologist. There I learned how to perform cell culture, subclone plasmid DNA and test whether a candidate gene was oncogenic. During college summers, I was fortunate enough to be able to go home and work in Periasamy Selvaraj's lab at Emory University. The 9 months of full time research I performed there were truly transformative. I learned how to think, how to read, and how to do experiments. Most importantly, I learned how to work. With a head filled with the stories of a younger Selva, I would work late into the night regularly performing western blots, purifying nanoparticles and performing mouse work. None of this could have been possible in the Selvaraj lab without the guidance of two amazing and caring graduate students Jaina Patel and Erica Bozeman. Each of them share with me a great deal of knowledge and provided me with friendship, despite the different stages of life we were in.

After college I spent two years at the National Institutes of Health's Vaccine Research Center (VRC) in the lab of Peter Kwong. There I worked primarily under Tongqing Zhou, who was the lead staff scientist in the lab and a brilliant crystallographer. Tongqing wasn't easy to please and almost never gave positive feedback. I always thought that this was his way of caring, of being a tough parent, and demanding more from me. Aside from Tongqing, I also received mentorship from the many postdoctoral and staff scientists in Peter's lab. Two particularly important mentors were Jason McLellan and Marie Pancera who still support me to this day. Beyond the scientific training I received doing protein crystallography at the VRC, one of the main things I walked away

with was a vision for how science could change the world. The discussions at the VRC were at the cutting edge and had a lofty vision for how the science we did could change the world.

My mentors at the University of Washington, Cole Trapnell and Jay Shendure have been phenomenal. I joined Cole's lab first in the summer of 2015. Prior to my rotation with Cole, I had just read about single cell genomics and I knew I wanted to be a part of the resolution revolution that was about to take place. Cole had just joined as a new faculty the summer before, and I remember this because the lab was empty when I started. At the time, my thinking in joining Cole's group was that I could learn from him, either through my work or osmosis, how one writes software for scientific computing. Although I feel no closer to this goal today, Cole allowed me to contribute my own strengths and natural talents to help further the group's broader research vision. The freedom, and trust that he has placed in me, allowed me to make the right mistakes and eventually find success. His enthusiasm and love for science are infectious and it is a pleasure to work with him everyday. Finally, Cole has a true appreciation for cool – an underrated quality in scientists. He knows how to tell a story and leave you wanting to be just like him. There are so many aspects of my current self that feel borrowed from Cole, but I don't think I'll ever give them back.

Although my interactions with Jay were limited for the first few years, I remember stealing inspiration in the form of little quotes I would hear from him or about him. Over time our relationship has deepened, but I still find myself looking forward to the next bit of inspiration or insight that I'm going glean from him. Apart from Jay's scientific brilliance, his kindness and dedication for treating people with respect are two of his most admirable qualities. This shows in the love he shows to trainees (both in his lab and outside of it) and to his family. Thanks for showing me that a scientific idol can also be a great person.

Finally, I want to thank the friends, colleagues, and mentors who supported me at the University of Washington. Thank you to members of the UW MST program and Genome Sciences department for creating an inspiring environment to do science. The wood sculpture of the double helix, and the sequences of the human and mouse genomes still inspire me to this day. Maybe one day, I'll

also be a part of such an endeavor.

Certain individuals made my training complete. Jonathan Packer took time to teach me so much of the command line magic and statistics I know today. His grasp of very difficult concepts and the ability to explain them in many different ways are admirable and formed the bulk of my hands-on computational training in the lab. Riza Daza taught me everything I know about sequencing and library preparation. Her persistence and thorough explanations helped some of the more difficult concepts in genomics get through. Jose McFaline sat right next to me and provided a lot of support both as a mentor and a friend. He has a way of always saying the right things, and has a heart of gold. His capacity to teach and help seems to know no bounds. Lauren Saunders has been one of my closest friends in the lab. It's hard to put a finger on why we get along so well, but if there's anyone I would want to spend a day doing molecular biology with, its her. Vijay Ramani was an early influence in my development as a thinker in genomics and I owe him a large thanks for being a good friend. More broadly, all the members of the Shendure and Trapnell labs, past and present have influenced the way I think and contributed in unknown ways to the successes I've experienced.

I also want to thank my committee members, Steven Henikoff, Stephen Tapscott, and Noah Simon for generously providing expertise, guidance, and perspective.

Over the course of completing a dissertation there's as much that happens outside the lab as inside the lab. Notably, Nick Hasle, Vanessa Gray, Greg Olson, Hugh Haddox, Una Nattermanns, Alex Ford, Ryan Volum, Sam Ainsley, Vineet Parkhe, Gwen Straley, Liza Severs and Matt Severs have made my time in Seattle absolutely sublime.

Finally, I thank my family for their infinite understanding, love and support. My parents always emphasized the importance of education, I think that enrolling in a decade long educational program was the only logical end to that thinking. Lastly, I want to thank my wife Katya Cherukumilli who has been a bedrock of support over the last decade. She's put up with a lot of late night and weekend shenanigans just so I could clone a plasmid one day faster or sequence an experiment

one day sooner. She's picked me up during my lows and helped bring me back to Earth during my highs. Her spirit, loving personality, and commitment to justice still serve as inspiration to me after all these years.

DEDICATION

This work is dedicated to my parents Srivatsan Ramachandran and Rajashree Srivatsan.

Amma, you have given me endless, unconditional love.
Appa, I hope you'll think of this work partly as your own, you've always been my number one fan, supporter and cheerleader.

Chapter 1: INTRODUCTION

1.1 The Sequencer as a Molecular Counter

In the biological sciences we are in the midst of a great cataloguing effort. With ever-increasing resolution and precision, we are compiling massive amounts of data. This is most apparent in genomics, where the cost of DNA sequencing has fallen precipitously and greatly exceeded the rate of decrease observed by Moore's law in computing (*1*). This rule of thumb was the observation that computing power doubled every 2 years for a fixed cost. This drop in computing costs has increased access to silicon-based technologies, which now impact nearly every sector, including transportation, telecommunication, warfare, business, consumer electronics, scientific research, and healthcare. If the biological sciences are to power a comparable paradigm shift, an important milestone will surely be the increased access to DNA sequencing.

The completion of the human genome [by the Human Genome Project (HGP) (*2*) and the Telomere-to-Telomere (T2T) Project (*3*, *4*)] and the genome sequences of other organisms spanning the tree of life (e.g. mouse (*5*), chimpanzee (*6*), zebrafish (*7*)) furnished a set of scaffolds for mapping short genomic segments. Having the genome sequence as a reference scaffold meant that instead of sequencing long fragments of DNA that needed to be stitched together, shorter "chopped up" fragments could be conveniently sequenced and queried against the reference genome scaffold(*8*). This simple idea has led to a renaissance in the development of new sequencing technologies and assays, which rely on using shorter DNA fragments as proxies for determining whether a particular DNA species was present (*9*). This thinking was quickly applied to study biology's central dogma, which resulted in the development of RNA-sequencing (*10*, *11*), ChIP-sequencing (*12*) and DNAse-seq (*13–15*). These assays, using short-read next generation DNA sequencing, respectively measure the production of RNAs, the binding of transcription factors to the genome, and the structure of chromatin. By biochemically converting each analyte into a sequencable frag-

ment that could be mapped to the reference genome, the DNA sequencer was repurposed as a generic molecular counter (*16*). Put another way, any observation – organic or inorganic – that could be linked to a change in the abundance of a DNA/RNA species, could be measured by DNA sequencing.

1.2 THE NEW QUANTITATIVE BIOLOGIST

Another consequence of sequencing genomes was that it laid bare the complexity of the biological algorithms contained within them. Prior to the completion of the human genome sequence, there was a widely held belief that it would explain how genes were regulated and elucidate the mutations that form the genetic basis for hereditary disease. This outlook was due in part, to a string of early successes, which seemed to indicate that specific inborn mutations could explain disease. Furthermore, many of these diseases (such as $/alpha$-, $/beta$-thalassemia, and sickle cell disease) exhibited a Mendelian pattern of inheritance – the disease state of an individual could be explained by the presence of dominant or recessive genetic sequences. Today, after sequencing hundreds of thousands of individuals, it is clear that the genetic and molecular basis for more common diseases such as Type 2 Diabetes Mellitus (T2DM) and Coronary Artery Disease (CAD) are not as straightforward. These studies indicate that mutations in a multitude of genes and in stretches of DNA sequence that regulate genes, each make small contributions to cumulative disease risk (*17*).

To illustrate this point, the models that best predict the risk of cardiovascular events are trained on large amounts of genetic data collected at the population scale (*18*). These quantitative models estimate a polygenic risk score (PRS) by tabulating a weighted sum of different mutations contained across many sites in the genome. This risk score can then be used to stratify patients with respect to risk of heart disease. Notably, PRS are agnostic to molecular mechanisms and act purely as a blackbox to predict risk.

The benefit of predicting a person's disease risk is self-evident. Individuals with increased risk for a disease (e.g. coronary artery disease, type 2 diabetes) could receive counseling, custom dietary plans, recommendations for lifestyle modification, and prophylactic pharmaceutical inter-

ventions. These interventions, all occurring prior to an individual's trip to a hospital, could inform lifestyle choices and dramatically decrease human suffering associated with disease. This vision of linking genotypic data with phenotypic data (health outcomes) is currently being realized through large national efforts to integrate the healthcare system and population-scale sequencing efforts.

For the task of predicting the risk of inheriting hereditary disease, the sequence of the genome is a useful abstraction; the inherited disease risk is materially propagated by the DNA that is stably inherited from generation to generation. Identifying the correct abstraction (or framing) for this problem has allowed for the unification of data collected asynchronously from human populations around the world – an effort that will have compounding rewards. Of note, the reference human genome is essential in this endeavor, because it provides the necessary scaffold. If other disciples are to emulate this effort, the importance of building the correct scaffold for data and information should be noted.

1.3 THE RIGHT LEVEL OF ABSTRACTION

Another grand vision in biology is the use of biological parts like tissues and cells in engineering applications. The modeling task underlying this vision is complex. Such a model would have to predict the behavior of a cell (or a group of cells) given an initial state and a set of inputs. However, finding the correct abstraction and representation of the cell is not intuitive. How does one describe what a cell is? One could pose the question, "What does a cell do?" Alternatively, there is an equally compelling line of questioning, "What attributes does a cell have? What molecules does it contain?"

These different frames can lead to wildly different representations (or encodings) of a cell. For example, if we were to take the route of describing the molecules contained within a cell, we could imagine modeling the cell at the atomic level. The cell would then be described as the set of molecular structures (atoms and their connectivities) and the locations of these structures within the cell. Correspondingly, a tissue would be a collection of these atomically-resolved cells, with interactions between the cells also resolved. Although this representation of the cell may

be the most faithful to material reality, physics-based atomic simulations of much less complex systems, (e.g. protein structures or small molecules) have proven difficult to model, because they become computationally intractable for larger atomic assemblies and over meaningful time scales. Furthermore, the training data needed to inform and refine such a model does not exist, due to the technical limitations of current measurement modalities – highly multiplexed imaging, at atomic resolution, in situ (within the cell).

Beyond the technical feasibility of an atomically resolved cell, it is unclear how this information would get encoded and aggregated. The insights from modeling genetic risk emphasize the importance of aggregating large amounts of data; a task that requires a unified scaffold. Second, such a model may be too complex for the cellular phenomena and behaviors that engineers and scientists are interested in harnessing. Building an atomic model of the cell would be akin to analysing the state of every transistor in a computer in order to reconstruct the operating system, programs, and data contained within it. Although this may be theoretically feasible, it would be the wrong way to approach the problem of figuring out what a computer is "doing."

1.4 THE GENOME, CELL AND DEVELOPMENT

Instead of simulating the atoms in the cell, a more practicable abstraction may be to consider the cell as the atomic unit. To draw this analogy out, atoms like Hydrogen and Helium have distinct chemical properties based on the number of protons contained within their nucleus. Analogously, we need comparable measurements of the cell that define certain cell classifications, and are linked to a cell's behavior.

To choose the measurement that would best describe the cell, cogitating a cell's properties is helpful. In the most simplistic form a cell is a container that compartmentalizes molecules and biochemical reactions, establishing an inside and an outside. This compartmentalization allows the cell to achieve functions that are essential for its own maintenance and self-preservation. Further, as cells replicate they acquire heritable changes and the cells that are most fit for their environment gain an advantage relative to other cells in the population. This description of the cell captures

the life cycle of prokaryotic cells (i.e. bacteria). To date bacteria have been thoroughly studied, effectively modeled, engineered, and deployed in the world.

To model the dynamics and behavior of cells or tissues contained within multicellular organisms (like humans), it is also important to consider how these collections of cells came to be. Most multicellular organisms result from the fusion of two gametes (the sperm and the egg in animals). Upon creation of this new organism, this primal cell rapidly divides, going through an orchestrated process of cellular differentiation. Over the course of this process, the organism takes form. The instruction set to specify this process is fully contained within the genome. At the molecular level, cascading sets of genes are differentially expressed, imbuing sets of cells with different attributes and abilities. This internal program unfolding within each cell, is then coordinated with and affected by surrounding cells. This process eventually leads to formation of a conscious and sentient organism capable of understanding the surrounding world, adapting to pathogens it encounters, and rebuilding itself upon injury.

1.5 Picking a measurement

In the execution of these dynamic programs, the genome plays a central role: it contains the blueprints to build cellular machinery, and it is the substrate upon which cellular computations are performed. In this cyclic loop, genes in the genome are expressed as RNA molecules, which are translated into proteins. Proteins act as the primary molecular machines present within the cell. They enable movement, facilitate cell division, sense the environment, and transduce the cell-relevant signals back to the genome for more computation. This property of proteins – the link between molecular and functional – means that the set of proteins expressed by a cell, could in principle, be a useful representation of that cell. Furthermore, properties include:

- The set of proteins is finite
- The same set of proteins is shared between cells
- A protein's presence and level of expression is tied to that cell's function

These properties make proteins an ideal candidate for measurement and representation of a cell. However, one missing component is the technical ability to make the measurement itself. To date few technologies have been able to faithfully measure the levels of all proteins in a single cell. When these measurements have been made, the scale of these experiments has been severely limited.

Fortunately the levels of messenger RNA (mRNA), the protein's molecular predecessor, correlate well with the proteins they code for (*19*). Moreover, mRNA molecules can be converted into sequence-able DNA fragments for quantification by next generation sequencing. This means that the ever dropping cost of sequencing can be brought to bear on profiling cells by sequencing the mRNA they express.

1.6 ISOLATING THE CELL

To measure the contents of a cell, the mRNA originating from each cell needs to be tallied. The earliest attempts to perform this assay involved the deposition of individual cells into individual wells of a microwell plate (*20*). Then, by preparing barcoded sequencing libraries from the mRNA in each well, the corresponding molecules from each cell could be sequenced and assembled into single cells. These early techniques including, MARS-seq (*21*), SMART-seq (*22*), demonstrated that single cell transcriptomes (the set of mRNA molecules present in a sample) could resolve individual cell types and characterize the cellular heterogeneity present within a collection of cells.

A further increase in the scale of these reactions was accomplished by first segregating cells from one another by encapsulating each cell in an emulsion (*23*). These protocols, such as dropSeq (*24*) and inDrops (*25*), co-encapsulated the cell with a bead containing barcoded reverse transcription (RT) primers. Because each bead contains a unique RT barcode, all the RNA molecules from a given cell would share a DNA-barcode sequence. The increase in scale achieved by these methods allowed for the measurement of many more single cell transcriptomes. This increase in scale allowed for the detection of rare and novel cell populations present in the sample.

1.7 COMBINATORIAL INDEXING

Unlike the methods mentioned before, single cell sequencing by combinatorial indexing (sci-seq), is a method for sequencing single cell transcriptomes without their prior isolation in wells or emulsions. Combinatorial indexing methods use a series of split-pool barcoding reactions to label co-transiting molecules. For single cell sequencing, the cell membrane or nuclear membrane act as the container that carries the DNA or RNA found within a cell, between barcoding reactions. The sci-seq paradigm has been adapted to capture a number of different molecular measurements including chromatin accessibility (sci-ATAC-seq) (*26*), gene expression (sci-RNA-seq) (*27*), methylation (sci-Met-seq) (*28*) and 3D genome conformation (sci-Hi-C) (*29*).

Furthermore, because sci-seq assays label single cells via their unique transit through barcoding reactions, increasing the number of barcoding reactions at each round increases the space of possible barcodes exponentially. The upshot is that more cells can be sequenced without incurring too many multiplets – two cells that have traveled through the same combination of wells. For example in 2 level sci-RNA-seq experiments, unique barcodes are acquired through an indexed RT reaction followed by an indexed PCR reaction (*27*). If 96 barcode sequences are used at each level of indexing, there are a total of 9,216 possible unique transits. Adding a third set of barcoding reactions (3-level sci-RNA-seq) with 96 additional indices increases the total barcode space to 884,736 possible combinations. This exponential scale that can be achieved by 3-level sci-RNA-seq in the number of barcodes has allowed for the processing of millions of cells from a single experiment (*30*).

The number of cells that can be profiled scales exponentially with the addition of more barcodes or levels of indexing. The sci-seq paradigm enables the profiling of millions of cells in a single experiment. While increasing the barcode space increases the number of single cells that can be resolved in a single experiment, there are limitations in the number of samples that can be processed simultaneously. Currently, each sample must be processed individually and loaded at an equal concentration into the first, barcoded, combinatorial indexing reaction. Using this method 10s to 100s of samples have been processed in parallel at great cost, labor, and experimental complexity.

1.8 SAMPLE MULTIPLEXING

Our ability to measure cells quickly exceeded our ability to make experimental contrasts with a single cell readout. To build accurate models of cellular systems, the data provided to train such a model must derive from diverse experimental conditions. To acquire this data, we effectively need ways to read out results from many perturbations with a highly resolved molecular readout. To tackle this problem, several groups have developed 'cellular hashing' methods, wherein cells from different samples are labeled and mixed prior to scRNA-seq (*31–33*). Although this greatly reduces per-sample costs, these hashing approaches require relatively expensive reagents (e.g. antibodies or chemically modified DNA oligos), use cell type-dependent protocols, and employ scRNA-seq platforms with a high per-cell cost. To address these shortcomings, and enable highly multiplexed experiments within a sci-seq framework I developed a hashing strategy (see **Chapter 2**) that uses unmodified oligonucleotides to label nuclei.

1.9 DEVELOPMENTAL ATLASES AND SPATIAL TRANSCRIPTOMICS

Sample multiplexing allows for the measurement of a cell's response consequent to chemical, genetic or other environmental perturbations. Although these perturbations may capture a cell's response to a drug or a genetic lesion, they don't provide insight into the full range of normal cellular states and transitions made by cells. To catalogue the complete set of cellular states, many single cell biologists have turned to study the developing organism. By documenting the set of cellular states and the transitions between them, the goal is to understand the molecular drivers underlying these cell state transitions.

To accomplish this feat, multiple groups have performed whole organism sequencing at single cell resolution (*27*, *30*, *34–37*), effectively profiling the diverse cellular populations and cellular states present over the course of development. These cellular states and the transitions between them, establish the natural progression of cells in an organism as it develops (*38*). However, thus far these datasets have been captured on disassociated or disaggregated cells. This means that there is no retention of spatial information in the assay. While there are emerging computational methods

and cutting edge technologies that attempt to address this problem, very few technologies spatially resolve the single cell transcriptome. As an extension of my work in sample multiplexing, in **Chapter 3** I describe the development of an assay called sci-Space, a spatial barcoding technology that is enabled by oligo hashing. Importantly, sci-Space can scale to the same extent as sci-RNA-seq.

1.10 CLOSING THOUGHT

To draw a final parallel to the genome projects, single cell biologists can currently profile cells in a shotgun manner akin to the effort of Celera Genomics and Craig Venter (*2*), however, the definitive scaffold need to interpret this data –the public human genome project – is still missing. The form of such a scaffold and the technologies used to construct it are being developed now.

Chapter 2: MASSIVELY MULTIPLEX CHEMICAL TRANSCRIPTOMICS AT SINGLE CELL RESOLUTION

Chapter 2 is adapted with minimal modification from:

Sanjay R Srivatsan*, José L. McFaline-Figueroa*, Vijay Ramani*, Lauren Saunders, Junyue Cao, Jonathan Packer, Hannah A. Pliner, Dana L. Jackson, Riza M. Daza, Lena Christiansen, Fan Zhang, Frank Steemers, Jay Shendure, Cole Trapnell (2020) Massively multiplex chemical transcriptomics at single cell resolution. *Science* 06 January 2020.

I contributed to all aspects of the paper, including the conception of the project, performance of the experiments, analysis of the data, and writing the initial manuscript draft.

2.1 ABSTRACT

High-throughput chemical screens typically employ coarse assays, e.g. cell survival, limiting what can be learned about mechanisms of action, off-target effects, and heterogeneous responses. Here we introduce sci-Plex, which uses 'nuclear hashing' to quantify global transcriptional responses to thousands of independent perturbations at single-cell resolution. As a proof-of-concept, we applied sci-Plex to screen 3 cancer cell lines exposed to 188 compounds. In total, we profiled 650,000 single-cell transcriptomes across 5,000 independent samples in one experiment. Our results reveal substantial intercellular heterogeneity in response to specific compounds, commonalities in response to families of compounds, and insight into differential properties within families. In particular, our results with HDAC inhibitors support the view that chromatin acts as an important reservoir of acetate in cancer cells.

2.2 INTRODUCTION

High-throughput screens (HTS) are a cornerstone of the pharmaceutical drug discovery pipeline (*39*, *40*). Conventional HTS have at least two major limitations. First, the readout of most HTS are restricted to gross cellular phenotypes, e.g. proliferation (*41*, *42*), morphology (*43*, *44*), or a

highly specific molecular readout (*45*, *46*). Subtle changes in cell state or gene expression, that might otherwise provide mechanistic insights or reveal off-target effects, are routinely missed.

Second, even when HTS are performed in conjunction with more comprehensive molecular phenotyping such as transcriptional profiling (*47–50*), a limitation of bulk assays is that even cells ostensibly of the same 'type' can exhibit heterogeneous responses (*51*, *52*). Such cellular heterogeneity can be highly relevant in vivo. For example, it remains largely unknown whether the rare subpopulations of cells that survive chemotherapeutics are doing so on the basis of their genetic background, epigenetic state, or some other aspect (*53*, *54*).

In principle, single-cell transcriptome sequencing (scRNA-seq) represents a form of high-content molecular phenotyping that could enable HTS to overcome both limitations. However, the per-sample and per-cell costs of most scRNA-seq technologies remain high, precluding even modestly sized screens. Recently, several groups developed 'cellular hashing' methods, wherein cells from different samples are molecularly labeled and mixed prior to scRNA-seq. However, current hashing approaches require relatively expensive reagents (e.g. antibodies (*55*) or chemically modified DNA oligos (*56*, *57*)), use cell type-dependent protocols (*58*), and/or employ scRNA-seq platforms with a high per-cell cost.

To enable cost-effective HTS with scRNA-seq-based phenotyping, we describe a novel sample labeling (hashing) strategy that relies on labeling nuclei with unmodified single-stranded DNA oligos. Recent improvements in single cell combinatorial indexing (sci-RNA-seq3) have lowered the cost of scRNA-seq library preparation to less than $0.01 per cell, with millions of cells profiled per experiment (*59*). Here we combine nuclear hashing and sci-RNA-seq into a single workflow for multiplex transcriptomics, termed 'sci-Plex'. As a proof-of-concept, we apply sci-Plex to perform a HTS of 3 cancer cell lines, profiling thousands of independent perturbations in a single experiment. We further explore how chemical transcriptomics at single cell resolution can shed light on mechanisms of action. Most notably, we find that gene regulatory changes consequent to treatment with HDAC inhibitors are consistent with the model that they interfere with proliferation by restricting a cell's ability to draw acetate from chromatin (*60*, *61*).

2.3 RESULTS

2.3.1 Nuclear hashing enables multi-sample sci-RNA-seq

Single-cell combinatorial indexing (sci-) methods use split-pool barcoding to uniquely label the molecular contents of large numbers of single cells or nuclei (*62*). Samples can be barcoded by these same indices, e.g. by placing each sample in its own well during reverse transcription in sci-RNA-seq (*59*, *63*), but such enzymatic labeling at the scale of thousands of samples is operationally infeasible and cost-prohibitive. To enable single-cell molecular profiling of a large number of independent samples within a single sci- experiment, we set out to develop a low-cost labeling procedure.

We noticed that single-stranded DNA (ssDNA) specifically stained the nuclei of permeabilized cells, but not intact cells **(Figure 3.1A; Figure 3.6A)**. We therefore postulated that a polyadenylated ssDNA oligonucleotide could be used to label populations of nuclei in a manner compatible with sci-RNA-seq **(Figure 3.1B; Figure 3.6B)**. To test this concept, we performed a 'barnyard' experiment. Human (HEK293T) and mouse cells (NIH3T3) were each separately seeded to 48 wells of a 96-well culture plate. We performed nuclear lysis in the presence of 96 well-specific polyadenylated ssDNA oligos ('hash oligos') and fixed the resulting nuclear suspensions with paraformaldehyde. Having labeled or 'hashed' the nuclei with a molecular barcode, we pooled nuclei and performed a 2-level sci-RNA-seq experiment. Because the hash oligos were polyadenylated, they had the potential to be combinatorially indexed identically to endogenous mRNAs. As intended, we recovered reads corresponding to both endogenous mRNAs (median 4,740 unique molecular identifiers (UMIs) per cell) and hash oligos (median 270 UMIs per cell).

We devised a statistical framework to identify the hash oligos associated with each cell at a frequency exceeding background. We observed 99.1% concordance between species assignments based on hash oligos vs. endogenous cellular transcriptomes **(Figure 3.1C; Figure 3.6C-F)**. Additionally, the association of hash oligos and nuclei was stable to a freeze-thaw cycle, highlighting the opportunity to label and store samples **(Figure 3.1D; Figure 3.6G,H)**. These results demonstrate that hash oligos stably label nuclei in a manner that is compatible with sci-RNA-seq.

In sci- experiments, 'collisions' are instances in which two or more cells are labeled with the same combination of barcodes by chance (*62*). To evaluate hashing as a means of detecting doublets resulting from collisions, we varied the number of nuclei loaded per PCR well, resulting in a range of predicted collision rates (7-23%) that was well matched by observation **(Figure 3.6I)**. Hash oligos facilitated the identification of the vast majority of interspecies doublets (95.5%) and otherwise undetectable within-species doublets **(Figure 3.1E; Figure 3.6J,K)**.

2.3.2 sci-Plex enables multiplex chemical transcriptomics at single-cell resolution

We next evaluated whether nuclear hashing could enable chemical screens, by labeling cells that had undergone a specific perturbation, followed by single-cell transcriptional profiling as a high-content phenotypic assay. We exposed A549, a human lung adenocarcinoma cell line, to one of four compounds dexamethasone (corticosteroid agonist), nutlin-3a (p53-Mdm2 antagonist), BMS-345541 (inhibitor of NF-κB-dependent transcription), or vorinostat/SAHA (histone deacetylase inhibitor) for 24 hours, across 7 doses in triplicate for a total of 84 drug/dose/replicate combinations and additional vehicle controls **(Figure 3.2A; Figure 3.7A)**. Nuclei from each well were labeled and subjected to sci-RNA-seq2 **(Figure 3.7B-D)**.

We used Monocle 3 (*59*) to visualize these data via Uniform Manifold Approximation and Projection (*64*) (UMAP) and louvain community detection to identify compound-specific clusters of cells, which were distributed in a dose-dependent manner **(Figure 3.2B,C; Figure 3.7E,F)**. To quantify the "population average" transcriptional response of A549 to each of the four drugs, we modeled each gene's expression as a function of dose through generalized linear regression. 7,561 genes were sensitive to at least one drug, and 3,189 genes were differentially expressed in response to multiple drugs **(Figure 3.8A)**. These included canonical targets of dexamethasone **(Figure 3.2D)** and nutlin-3a **(Figure 3.2E)**. Gene ontology analysis of differentially expressed genes revealed the involvement of drug-specific pathways (e.g. hormone signaling for dexamethasone; p53 signaling for nutlin-3a; **Figure 3.8B**). Additionally, we evaluated whether the number of cells recovered at each concentration could be used to infer toxicity akin to traditional screens. After fitting a

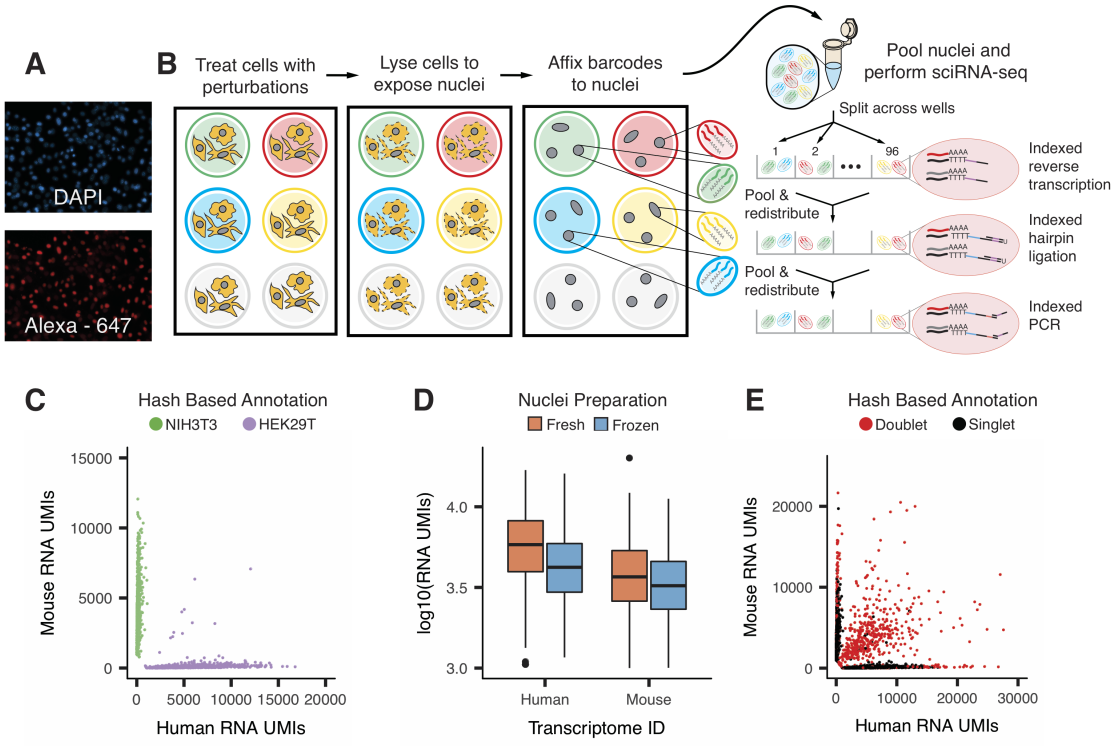

Figure 2.1: **sci-Plex uses polyadenylated single stranded oligonucleotides to label nuclei, enabling cell hashing and doublet detection.** **(A)** Fluorescent images of permeabilized nuclei after incubation with DAPI (top) and an Alexa-647 conjugated single stranded oligonucleotide (bottom). **(B)** Overview of sci-Plex. Cells corresponding to different perturbations are lysed in-well, their nuclei labeled with well-specific "hash" oligos, followed by fixation, pooling and sci-RNA-seq. **(C)** Scatter plot depicting the number of unique molecular identifiers (UMIs) from single cell transcriptomes derived from a mixture of hashed human HEK293T cells and murine NIH3T3 cells. Points colored based on hash oligo assignment. **(D)** Boxplot depicting the number of mRNA UMIs recovered per cell for fresh vs. frozen human and mouse cell lines. **(E)** Scatter plot of overloading experiment; axes as in panel **C**. Identified Hash oligo collisions (red) identify cellular collisions with high sensitivity.

response curve to the recovered cellular counts, we inferred a 'viability score' from sci-Plex data, a metric which was concordant with gold standard measurements **(Figure 3.2F;Figure 3.7G-I)**.

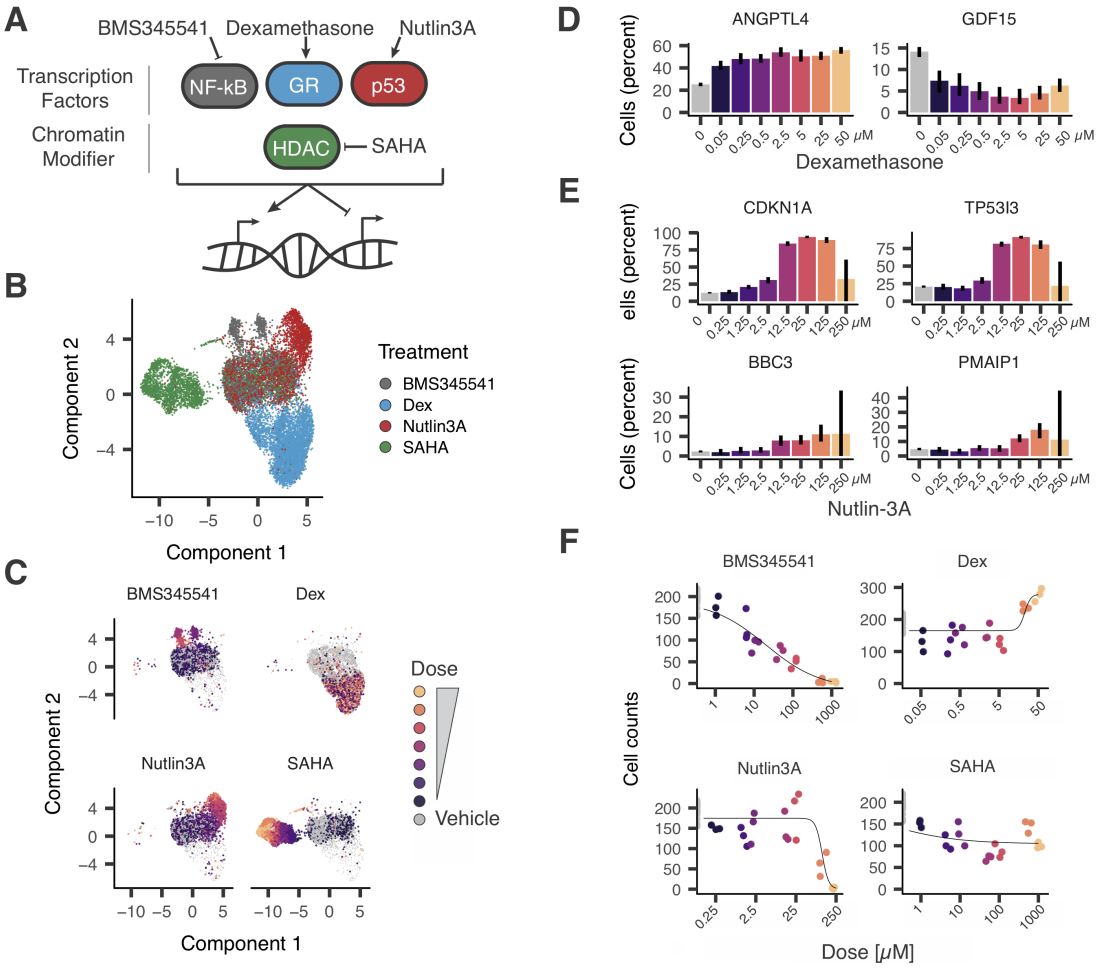

Figure 2.2: **sci-Plex enables multiplex chemical transcriptomics at single cell resolution.** **(A)** Cartoon representation depicting compounds and corresponding targets assayed within the pilot sci-Plex experiment. A549 lung adenocarcinoma cells were treated with either vehicle (DMSO or ethanol) or one of four compounds (BMS345541, dexamethasone, nutlin-3a or SAHA). **(B)** UMAP embedding of chemically perturbed A549 cells colored by drug treatment. **(C)** UMAP embedding of chemically perturbed A549 cells faceted by treatment with cells colored by dose. **(D,E)** Expression of a canonical **(D)** glucocorticoid receptor activated (*ANGPTL4*) and repressed (*GDF15*) target genes as a function of dexamethasone dose, or **(E)** p53 target genes as a function of nutlin-3a dose. Y-axes indicate percentage of cells with at least one read corresponding to the transcript. **(F)** Dose-response viability estimates for BMS345541, dexamethasone, nutlin-3a and SAHA-treated A549 cells, based on the relative number of cells recovered at each dose.

2.3.3 sci-Plex scales to thousands of samples and enables HTS

To assess how sci-Plex scales for HTS, we performed a screen of 188 compounds targeting a diverse range of enzymes and molecular pathways **(Figure 3.3A)**. Half of this panel was chosen to target transcriptional and epigenetic regulators. The other half was chosen to sample diverse mechanisms of action. We exposed three well-characterized human cancer cell lines (A549 (lung adenocarcinoma), K562 (chronic myelogenous leukemia), MCF7 (mammary adenocarcinoma)) to each of these 188 compounds at four doses (10 nM, 100 nM, 1μ M, 10 μM) in duplicate, randomizing compounds and doses across well positions in replicate culture plates. These conditions, together with vehicle controls, accounted for 4,608 of 4,992 independently treated cell populations in this experiment. After treatment we lysed cells to expose nuclei, hashed them with a unique combination of two oligos **(Figure 3.9A)**, and performed sci-RNA-seq3 (*59*). After sequencing and filtering based on hash purity **(Figure 3.9B-F)**, we obtained transcriptomes for 649,340 single cells, with median mRNA UMI counts of 1,271, 1,071, and 2,407 for A549, K562 and MCF7, respectively **(Figure 3.10A)**. The aggregate expression profiles for each cell type were highly concordant between replicate wells (Pearson correlation = 0.99) **(Figure 3.10B)**.

Visualizing sci-RNA-seq profiles separately for each cell line revealed compound-specific transcriptional responses and patterns that were common to multiple compounds. For each of the cell lines, UMAP projected most cells into a central mass, flanked by smaller clusters **(Figure 3.3B)**. These smaller clusters were largely comprised of cells treated with compounds from only one or two compound classes **(Figure 3.11; Figure 3.12A-C)**. For example, A549 cells treated with triamcinolone acetonide, a synthetic GR agonist, were markedly enriched in one such small cluster, comprising 95% of its cells (Fisher's exact test, FDR $<$ 1%, **Figure 3.12D,E**). Although many drugs were associated with a seemingly homogenous transcriptional response, we also identified cases in which distinct transcriptional states were induced by the same drug. For example, in A549, the microtubule stabilizing compounds epothilone A and epothilone B were associated with three such focal enrichments, each comprised of cells from both compounds at all 4 doses **(Figure 3.12F,G)**. The cells in each focus were distinct from one another, but transcriptionally similar to

other treatments – either a recently identified microtubule destabilizer, rigosertib (*65*), the SETD8 inhibitor UNC0397, or untreated proliferating cells **(Figure 3.12H)**.

We next assessed the effects of each drug on the 'population average' transcriptome of each cell line. In total, 6,238 genes were differentially expressed in a dose-dependent manner in at least one cell line (FDR $< 5\%$; **Figure 3.13**;). Bulk RNA-seq measurements collected for 5 compounds, across 4 doses and vehicle agreed with averaged gene expression values and estimated effect sizes across identically treated single cells, although correlations between small effect sizes were diminished **(Figure 3.14)**. Moreover, sci-Plex dose-dependent effect profiles correlated with compound matched L1000 measurements (*49*) **(Figure 3.15)**.

Genes associated with the cell cycle were highly variable across individual cells, and many drugs reduced the fraction of cells that expressed proliferation marker genes **(Figure 3.15-Figure 3.17)**. In principle, scRNA-seq should be able to distinguish shifts in the proportion of cells in distinct transcriptional states from gene regulatory changes within those states. In contrast, bulk transcriptome profiling would confound these two signals **(Figure 3.18A)** (*52*). We therefore tested for dose-dependent differential expression on subsets of cells corresponding to the same drug but expressing high vs. low levels of proliferation marker genes **(Figure 3.18B)**. Correlation between the dose-dependent effects on the two fractions of each cell type varied across drug classes (Figure 3.18C), with some frankly discordant effects for individual compounds **(Figure 3.18D)**. Viability analysis, performed as in the pilot experiment, revealed that after drug exposure at the highest dose, only 52 (27%) compounds caused a drop in viability of 50% or more **(Figure 3.3C; Figure 3.10C)**. Amongst the drugs that reduced viability, we observed a higher sensitivity of K562 to the Src/Abl inhibitor bosutinib **(Figure 3.3C)**, a result we confirmed via cell counting **(Figure 3.19A)**. This result is consistent with K562 cells harboring a constitutively active BCR-ABL fusion kinase (*66*) and an observed increased sensitivity of hematopoietic and lymphoid cancer cell lines to Abl inhibitors (*67*) **(Figure 3.19B)**.

To assess whether each compound elicited similar responses across the three cell lines, we clustered compounds using the effect sizes for dose-dependent genes as loadings in each cell line **(Figure 3.20-Figure 3.35)**. Joint analysis of the three cell lines revealed common and cell type

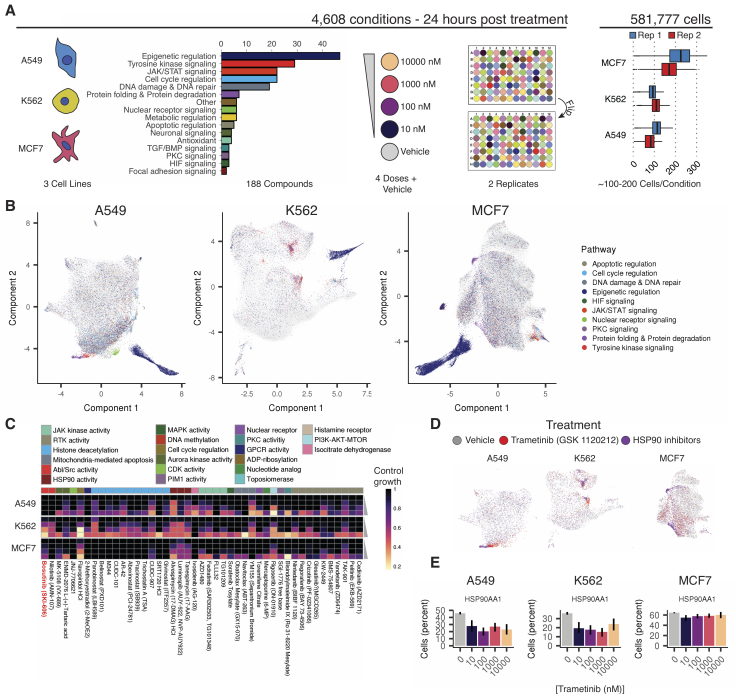

Figure 2.3: **sci-Plex enables global transcriptional profiling of thousands of chemical perturbations in a single experiment. (A)** Schematic of the large-scale sci-Plex experiment (sci-RNA-seq3). 188 small molecules were tested for their effects on A549, K562 and MCF7 human cell lines, each at 4 doses and in biological replicate, after 24 hours of treatment. The plate positions of doses and drugs were varied between replicates and a median of 100 to 200 cells were recovered per condition. **(B)** UMAP embeddings of A549, K562 and MCF7 cells in our screen with each cell colored by the pathway targeted by the compound to which a given cell was exposed. To facilitate visualization of significant molecular phenotypes, we added transparency to cells treated with compound/dose combinations that did not appreciably alter the corresponding cells' distribution in UMAP space as compared to vehicle controls (Fisher's exact test, FDR $< 1\%$). **(C)** Viability estimates obtained from hash-based counts of nuclei at each dose of selected compounds (bosutinib highlighted in red text). Rows represent compound doses increasing from top to bottom and columns represent individual compounds. Annotation bar at top depicts the broad cellular activity targeted by each compound. **(D)** UMAP embeddings highlighted by treatment with the MEK inhibitor trametinib (red), an HSP90 inhibitor (purple), or vehicle control (gray). **(E)** *HSP90AA1* expression levels in cells exposed to increasing doses of trametinib. Y-axes indicate percentage of cells with at least one read corresponding to the transcript.

specific responses to different compounds **(Figure 3.36-Figure 3.37)**. For example, trametinib, a MEK inhibitor, induced a transcriptionally distinct response in MCF7. Inspection of UMAP pro-

jections revealed trametinib-treated MCF7 interspersed amongst vehicle controls, reflecting limited effects. In contrast, trametinib-treated A549 and K562 cells, which harbor activating KRAS and ABL mutations (*68*), respectively, were tightly clustered, consistent with a strong, specific transcriptional response to inhibition of MEK signaling by trametenib **(Figure 3.3D)**. Further, we observed that these A549 and K562 cells appeared proximal to clusters enriched with inhibitors of HSP90, a key chaperone for protein folding **(Figure 3.3D)**. This observation was corroborated by concordant changes in *HSP90AA1* expression in Trametinib-treated cells **(Figure 3.3E)**. Analysis of Connectivity Map data (*49*, *50*) revealed further evidence that MEK inhibitors do indeed induce highly similar gene expression signatures to HSP90 perturbations **(Figure 3.19C)** especially in A549 but not in MCF7 **(Figure 3.19D-E)**. These results are concordant with previous observations of the regulation of *HSP90AA1* downstream of MEK signaling (*69*) and suggests that similarity in single-cell transcriptomes treated with distinct compounds can highlight drugs that target convergent molecular pathways.

2.3.4 Inference of chemical and mechanistic properties of HDAC inhibitors

For each of the three cell lines, the most prominent compound response was comprised of cells treated with one of seventeen histone deacetylase (HDAC) inhibitors (dark blue in **Figure 3.3B**;). To assess the similarity of the dose-response trajectories between cell lines, we aligned HDAC-treated cells and vehicle-treated cells from all three cell lines using a mutual-nearest neighbor (MNN) matching approach (*70*) to produce a consensus HDACi trajectory, which we term 'pseudo-dose' (analogous to pseudotime (*71*)) **(Figure 3.4B; Figure 3.38)**. We observed that some HDAC inhibitors induced homogeneous responses, with nearly all cells localized to a relatively narrow range of the HDACi trajectory at each dose (e.g. pracinostat in A549), while other drugs induced much greater cellular heterogeneity **(Figure 3.4B; Figure 3.39)**.

Such heterogeneity could be explained by cells executing a defined transcriptional program asynchronously, with the dose of drug that the cells are exposed to modulating the rates of their progression through it. To test this hypothesis, we sequenced the transcriptomes of 64,440 A549

cells that were treated for 72 hours with one of 48 compounds, including many of the HDAC inhibitors from the large sci-Plex screen. Upon accounting for confluency dependent cell cycle effects and MNN alignment **(Figure 3.40-Figure 3.42)**, the co-embedded UMAP projection revealed new focal concentrations of cells at 72 hours that were not evident at the 24 hour time point, e.g. SRT1024 **(Figure 3.43)**. However, for the majority of HDAC inhibitors tested, we did not observe that cells at a given dose moved farther along an aligned HDAC trajectory at 72 hours **(Figure 3.44)**. This suggests that the dose of many HDAC inhibitors governs the magnitude of a cell's response rather than its rate of progression and that any observed heterogeneity cannot be attributed solely to asynchrony **(Figure 3.44)**.

Next, we assessed whether a given HDAC inhibitor's target affinity explained its global transcriptional response to the compound. We used dose-response models to estimate each compound's 'transcriptional EC50 (TC50)', i.e. the concentration needed to drive a cell halfway across the HDACi pseudodose trajectory **(Figure 3.45A)**. To compare the transcriptionally-derived measures of potency with the biochemical properties of each compound, we collected published IC50 values for each compound from in vitro assays performed on 8 purified HDAC isoforms. With the exception of 2 relatively insoluble compounds, our calculated TC50 values increased as a function of compound IC50 values **(Figure 4C; Figure 3.45B-C)**.

To assess the components of the HDAC inhibitor trajectory, we performed differential expression analysis using pseudodose as a continuous covariate. Of the 4,308 genes that were significantly differentially expressed over this consensus trajectory, 2,081 (48%) responded in a cell type-dependent manner, while 942 (22%) exhibited the same pattern in all three cell lines (Figure 3.46A,B). One prominent pattern shared by the three cell lines was an enrichment for genes and pathways indicative of progression towards cell cycle arrest **(Figure 3.46C; Figure 3.47A,B)**. DNA content staining and flow cytometry confirmed that HDAC inhibition resulted in the accumulation of cells in the G2/M phase of the cell cycle (*72*) **(Figure 3.47C,D)**.

The shared response to HDAC inhibition included not only cell cycle arrest, but also the altered expression of genes involved in cellular metabolism **(Figure 3.46C)**. Histone acetyltransferases and deacetylases regulate chromatin accessibility and transcription factor activity through the ad-

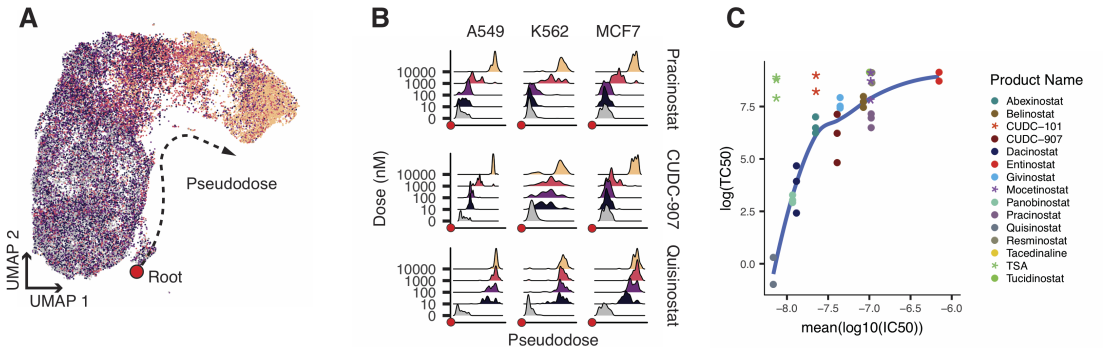

Figure 2.4: **HDAC inhibitor trajectory captures cellular heterogeneity in drug response and biochemical affinity.** **(A)** Nearest-neighbor alignment and UMAP embedding of transcriptional profiles of cells treated with one of seventeen HDAC inhibitors. Pseudodose root displayed as a red dot. **(B)** Ridge plots displaying the distribution of cells along pseudodose by dose. Shown for 3 HDAC inhibitors with varying biochemical affinities. **(C)** Relationship between log transcriptional EC50 (TC50) and average log10(IC50) from in vitro measurements. Asterisks indicate compounds with a solubility below 200 mM (in DMSO) that were not included in the fit.

dition or removal of charged acetyl groups (*73–75*). Acetate, the product of HDAC class I-,II- and IV-mediated histone deacetylation and a precursor to acetyl-CoA, is required for histone acetylation but also has important roles in metabolic homeostasis (*61*, *76*, *77*). Inhibition of nuclear deacetylation limits recycling of chromatin-bound acetyl groups for both catabolic and anabolic processes (*77*). Accordingly, we observed that HDAC inhibition led to sequestration of acetate in the form of markedly increased acetylated lysine levels after exposure to a 10μM dose of the HDAC inhibitors pracinostat and abexinostat **(Figure 3.48)**.

Upon further inspection of pseudodose-dependent genes, we observed that enzymes critical for cytoplasmic acetyl-CoA synthesis from either citrate (*ACLY*) or acetate (*ACSS2*) were upregulated (Figure 3.5A). Genes involved in cytoplasmic citrate homeostasis (*GLS*, *IDH1*, and *ACO1*), citrate cellular import (*SLC13A3*) and mitochondrial citrate production and export (*CS*, *SLC25A1*) were also upregulated. Upregulation of *SIRT2*, which deacetylates tubulin, was also observed in response to HDAC inhibition.

Together with increases in chromatin-bound acetate, these transcriptional responses suggest

a metabolically consequential depletion of cellular acetyl-CoA reserves in HDAC-inhibited cells **(Figure 3.5B)**. To validate this further, we sought to shift the distribution of cells along the HDAC inhibitor trajectory by modulating cellular acetyl-CoA levels. We treated A549 and MCF7 cells with pracinostat in the presence and absence of acetyl-CoA precursors (acetate, pyruvate or citrate) or inhibitors to enzymes (*ACLY*, *ACSS2*, or *PDH*) involved in replenishing acetyl-CoA pools. After treatment, cells were harvested and processed via sci-Plex and trajectories constructed for each cell line **(Figure 3.49-Figure 3.50)**. In both A549 and MCF7, acetate, pyruvate and citrate supplementation were capable of blocking pracinostat-treated cells from reaching the end of the HDACi trajectory **(Figure 3.49F,J,H,L)**. In MCF7, both *ACLY* and *ACSS2* inhibition shifted cells further along the HDACi trajectory, although no such shift was observed in A549 **(Figure 3.49G,K,I,M)**. Taken together, these results suggest that a major feature of the response of cells to HDAC inhibitors, and possibly their associated toxicity, is the induction of an acetyl-CoA deprived state.

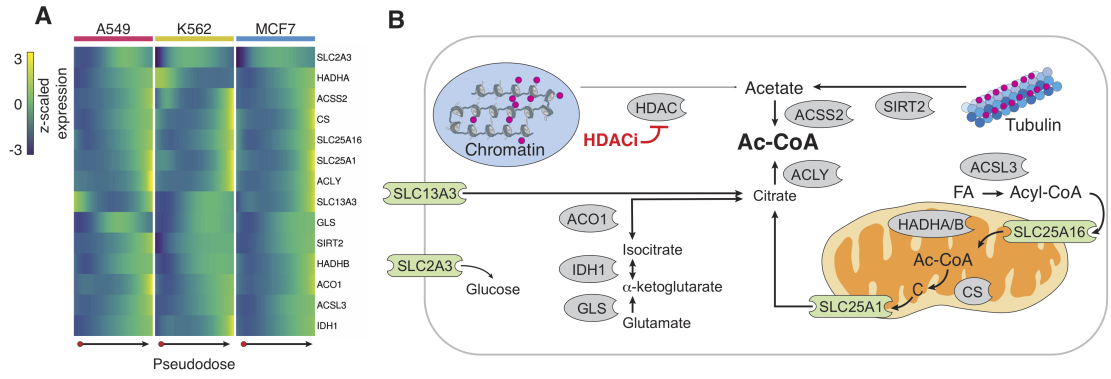

Figure 2.5: **HDAC inhibitors shared transcriptional response indicative of acetyl CoA deprivation. (A)** Heatmap of row-centered and z-scaled gene expression depicting the upregulation of pseudodose-dependent genes involved in cellular carbon metabolism. **(B)** Cartoon of the roles of genes from panel **A** in cytoplasmic acetyl-CoA regulation. Red circles indicate acetyl groups. Enzymes shown in gray. Transporters shown in green (FA = fatty acid; Ac-CoA = Acetyl-CoA; C = Citrate).

2.4 DISCUSSION

Here we present sci-Plex, a massively multiplex platform for single-cell transcriptomics. sci-Plex uses chemical fixation to cost-effectively and irreversibly label nuclei with short, unmodified single-stranded DNA oligos. In the proof-of-concept experiment described here, we applied sci-Plex to quantify the dose-dependent responses of cancer cells to 188 compounds through an assay that is both high content (global transcription) and high resolution (single cell). By profiling several distinct cancer cell lines, we distinguished between shared and cell line-specific molecular responses to each compound.

sci-Plex offers some unique advantages over conventional HTS: sci-Plex can distinguish a compound's distinct effects on cellular subsets (including complex in vitro systems e.g. cellular reprogramming, organoids, synthetic embryos); it can unmask heterogeneity in cellular response to a perturbation; and it can measure how drugs shift the relative proportions of transcriptionally distinct subsets of cells. Highlighting these features, our study provides insight into the mechanism of action of HDAC inhibitors. Specifically, we find that the main transcriptional responses to HDAC inhibitors involve cell cycle arrest and marked shifts in genes related to acetyl-CoA metabolism. For some HDAC inhibitors, we observe clear heterogeneity in responses observed at the single cell level. Although HDAC inhibition is conventionally thought to act through mechanisms directly involving chromatin regulation, our data support an alternative model, albeit not a mutually exclusive one, wherein HDAC inhibitors impair growth and proliferation by interfering with a cancer cell's ability to draw acetate from chromatin (*60*, *61*, *77*). As such, variation in cells' acetate reservoirs is a potential explanation for their heterogeneous responses to HDAC inhibitors.

As the cost of single cell sequencing continues to fall, the opportunities for leveraging sci-Plex for basic and applied goals in biomedicine may be substantial. The proof-of-concept experiments described here – nearly 5,000 independent treatments, transcriptional profiling of over 100 single cells per treatment – can potentially be scaled towards a comprehensive, high-resolution atlas of cellular responses to pharmacologic perturbations (*e.g.* hundreds of cell lines or genetic backgrounds, thousands of compounds, multi-channel single cell profiling, etc.). The ease and low cost

of oligo hashing, coupled with the flexibility and exponential scalability of single cell combinatorial indexing, would facilitate this goal.

2.5 Acknowledgments

We thank members of the Shendure lab, Trapnell lab and others, particularly A. Adey, G. Booth, A. Hill, S. Henikoff, K. Cherukumilli, P. Selvaraj, and T. Zhou for helpful suggestions, discussion and mentorship. D. Prunkard and A. Leith for assistance in flow sorting. Funding: This work was funded by grants from the NIH (DP1HG007811 and R01HG006283 to J.S.; DP2 HD088158 to C.T.), the W. M. Keck Foundation (to C.T. and J.S.), the NSF (DGE-1258485 to S.S). The Paul G. Allen Frontiers Group (to J.S. and C.T.). J.S. is an Investigator of the Howard Hughes Medical Institute. L.C. , FZ, and F.J.S. declare competing financial interests in the form of stock ownership and paid employment by Illumina, Inc. Author contributions: S.R.S., J.L.M., V.R., J.S., C.T. conceived the project; S.R.S., J.L.M., V.R. designed experiments; F.Z, L.C, F.S provided reagents and performed sequencing; S.R.S., J.L.M., C.T. analyzed the data; S.R.S., J.L.M., V.R., J.S., C.T. wrote the manuscript. Competing interests: One or more embodiments of one or more patents and patent applications filed by Illumina and University of Washington may encompass the methods, reagents, and data disclosed in this manuscript. Data and materials availability: Processed and raw data can be downloaded from NCBI GEO (#GSE139944). Code used to perform analyses can be accessed on Zenodo and https://github.com/cole-trapnell-lab/sci-plex.

2.6 Materials and Methods

2.6.1 Cell Culture

A549 cells and K562 cells were a kind gift from Dr. Robert Bradley (UW) and Dr. David Hawkins (UW), respectively. MCF7 (cat no. HTB-22), NIH3T3 (cat no. CRL-1658) and HEK293T (cat no. CRL-11268) cells were purchased from ATCC. A549 and MCF7 cells were cultured in DMEM (ThermoFisher, 11995073) media supplemented with 10% FBS (ThermoFisher, cat no. 26140079) and 1% penicillin-streptomycin (ThermoFisher, 15140122). K562 cells were cultured in RPMI

1640 (Fisher Scientific, cat no. 11-875-119) supplemented with 10% FBS and 1% penicillin-streptomycin and maintained between 0.2-1 x 106 cells/ml. All cells were cultured at 37C with 5% CO2. Adherents cells were split when they reached 90% confluence by washing with DPBS (Life Technologies, cat no. 14190-250), trypsinizing using TryPLE (Fisher Scientific, cat no. 12-604-039) and split at either 1:4 (MCF7) or 1:10 (A549, NIH3T3 and HEK293T).

2.6.2 Compound Preparation

Dexamethasone was purchased from Sigma-Aldrich and resuspended in molecular biology grade ethanol (Fisher Scientific). BMS-345541 (S8044), Vorinostat (S1047), and Nutlin-3a (S8059) were acquired from Selleck Chemicals and resuspended in DMSO (VWR Scientific, 97063-136). Cherry-picked 96-well compound screens were acquired from Selleck Chemicals resuspended to 10 mM in DMSO (Table S2). Compounds were diluted in their respective vehicle to 1000x of their desired treatment concentration and stored at -80C until use.

2.6.3 Drug treatment

For 96-well experiments, adherent cells were trypsinized, washed with PBS and plated in tissue culture treated 96 well flat bottom plates $ThermoFisherScientific, catno.12-656-66$ at 25,000 cells per well in 100 μL of media. Suspension cells were washed with PBS and plated in 96 well V-bottom tissue culture plates (Thermo Fisher Scientific, cat no. 549935) at 25,000 cells per well in 100 μL of media. Cells were allowed to recover for 24 hours before treatment with 1 μL of a 1:10 dilution of the appropriate compound or vehicle in PBS to maintain a vehicle concentration of 0.1% for all wells. Cells were then exposed to small molecules at the specified concentration for either 24 or 72 hours. For experiments where cells were co-treated with HDAC inhibitors and either acetate, pyruvate, citrate, ACSS2 inhibitor, ACLY inhibitor or PDH inhibitor, cells were treated 24 hours after plating and harvested after 24 hours. In this set of experiments, all wells contained a final concentration of 0.2% DMSO to match treatment with both the HDAC inhibitor and inhibitors of metabolic processes.

2.6.4 CellTiter Glo

A549, MCF7 and K562 cells were seeded in 96 well plates, allowed to attach for 24 hours and treated with BMS345541, dexamethasone, nutlin-3A, SAHA, as described above. 24 hours post treatment, plates were allowed to reach room temperature and viability estimated using the CellTiter-Glo viability assay (Promega) according to manufacturer's instructions. Luminescence was recorded using a BioTek synergy plate reader. For each drug treatment luminescence readings were normalized to the average luminescence intensities of vehicle DMSO treated wells.

2.6.5 Cell counts of bosutinib exposed cells

A549, MCF7 and K562 cells were seeded in 12 well plates at 2.8 x 105 cells per well. After 24 hours to allow for A549 and MCF7 attachment, cells were exposed for 24 hours to 0.1, 1 and 10 μM bosutinib or DMSO vehicle control. After treatment, adherent cells were detached using TrypLE or directly resuspended in 1 mL of media and cells counted on a Countess II FL automated cell counter (ThermoFisher).

2.6.6 Cancer cell line encyclopedia and connectivity map data and analysis

Pharmacological profiling data was downloaded from the Cancer cell line encyclopedia (CCLE) data portal (https://portals.broadinstitute.org/ccle/data). Data was the isolated and plotted for cell line of haematopoietic and lymphoid, lung and breast tissue origin exposed to the Abl inhibitors AZD0530 and nilotinib. Connectivity map (CMAP) data was downloaded from the CLUE command app in the CMAP data portal. Top connections and connectivity scores were exported between the MEK inhibitor perturbagen class and HSP inhibitor perturbagen class across all cell lines (Summary) or individual cell lines that overlap with our study (A549 and MCF7). Results were then filtered for data from inhibitor exposure. To determine how connectivities change across all vs. individual cell lines, we filtered for the top connections that overlap with the connectivity summary in data from individual cell lines. Connectivity scores were subjected to a threshold value of 90 as in the associated CMAP study (*49*).

2.6.7 Flow cytometry

A549 and MCF7 cells were seeded in 6 cm dishes at 1.6 x 106 cells per plate. K562 cells were seeded in T25 cm2 flasks at 1.6 x 106 cells per flask. After 24 hours to allow for A549 and MCF7 attachment cells were exposed for 24 hours to 10 μM abexinostat, 10 μM pracinostat or DMSO as a vehicle control. After treatment cells were harvested as described above, pellets washed twice in PBS, resuspended in 500 μL of cold PBS and fixed by the addition of 5 mL of ice-cold ethanol while vortexing at low speed. Cells were stored at -20C prior to processing for flow cytometry analysis. For flow cytometry, ethanol was removed and fixed cells washed twice with PBS containing 1% BSA (PBS-B) and blocked for 1 hour at room temperature. Then, blocking buffer was removed and cells were incubated in PBS containing 1% BSA and 0.1% tryton X-100 (PBS-BT) as well as a 1:500 dilution of mouse anti-acetyl-lysine antibody (cat no. ICP0390, ImmuneChem Pharmaceuticals Inc) for 2 hours at room temperature. After incubation, cells were washed twice with PBS-BT and incubated with goat anti-mouse Alexa-647 in PBS-BT for 1 hour at room temperature. Lastly, cells were washed twice with PBS-BT, once with PBS-B and resuspended in PBS-B containing 5 μg/ml Hoechst 33258 (Life Sciences Technologies) to stain the DNA. Then the levels of total acetylated-lysine and DNA content was analyzed by flow cytometry on an LSRII flow cytometer (BD Biosciences). Quantification and downstream analysis was performed using FlowJo10 (FlowJo.LLC).

2.6.8 Cell harvest, nuclei isolation and sample hashing

For the harvest of adherent cells, media was removed, and cells were rinsed with 100 μL of DPBS and tryspinized with 50 μL of Tryp-LE for 15 minutes at 37C. Once cells had detached from the culture plate, the reaction was quenched with 150 μL of ice-cold DMEM containing 10% FBS. Cell suspensions were generated by pipetting and the entire volume was transferred to a 96 well V-bottom plate. Cells were then pelleted by centrifugation at 300 x g for 6 minutes, washed with 100 μL of ice-cold DPBS and re-pelleted at 300 x g for 6 minutes.

Lysis was conducted in the 96 well V-bottom plate. Following removal of PBS, cell suspensions

were lysed and labeled with 50 μL of cold lysis buffer (10 mM Tris-HCl, pH 7.4, 10 mM NaCl, 3 mM MgCl2, 0.1% IGEPAL CA-630) (*62*) supplemented with 1% Superase RNA Inhibitor and 400 femtomoles of hashing oligo of the form 5'-GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAG-[10bp-barcode]-BAAAAAAAAAAAAAAAAAAAAAAAAAAAAA-3' where B is G, C or T (IDT). For the large compound screen, 500 femtomoles of an additional oligo was used to uniquely index each 96 well treatment plate. After lysis with 3 strokes of multichannel pipette, cells were fixed by addition of 200 μL of fixation buffer (5% Paraformaldehyde, 1.25x PBS). Nuclei were then fixed on ice for 15 minutes before pooling into a trough. Nuclei were pooled by plate into a 50 mL conical tube and pelleted by centrifugation at 500 x g for 5 minutes. Subsequently, cells were resuspended in 500 μL of nuclei suspension buffer (NSB; (10 mM Tris-HCl, pH 7.4, 10 mM NaCl, 3 mM MgCl2, 1% Superase RNA Inhibitor, 1% 0.2mg/mL Ultrapure BSA)). Finally, nuclei from all plates were pooled into a single conical tube and nuclei were pelleted by centrifugation at 500 x g for 5 minutes. Nuclei were then resuspended in 1mL of NSB and flash frozen into liquid nitrogen in 100 μL aliquots. Nuclei were then stored at -80C until further processing with sci-RNA-seq.

2.6.9 Preparation of sci-RNA-seq2 libraries

Frozen nuclei were thawed over ice and spun down at 500g for 5 minutes. Cells were then permeabilized in permeabilization buffer (NSB + 0.25% Triton-X) for 3 minutes and then spun down. Following another a wash in NSB, two-level sci-RNA-seq libraries prepared as previously described (*63*). Briefly, nuclei were pelleted at 500 x g for 5 minutes, and resuspended in 100 μL of NSB. Cell counts were obtained by staining nuclei with 0.4 % trypan blue (Sigma-Aldrich) and counted using a hemocytometer. 5000 nuclei in 2 μL of NSB and 0.25 μL of 10 mM dNTP mix (Thermo Fisher Scientific, cat no. R0193) were then distributed onto a skirted twin.tec 96 well LoBind plate (Fisher Scientific, cat no. 0030129512) after which 1 μL of uniquely indexed oligo-dT (25 μM)(*63*) was added to every well, incubated at 55C for 5 minutes and placed on ice. 1.75 μL of reverse transcription mix (1μL of Superscript IV first-strand buffer, 0.25 μL of 100 mM

DTT, 0.25 μL of Superscript IV and 0.25 μL of RNAseOUT recombinant ribonuclease inhibitor) was then added to every well and plates incubated at 55C for 10 minutes and placed on ice. 5 μL of stop solution (40 mM EDTA, 1 mM spermidine and 0.5% BSA) were added to each well to stop the reaction. Wells were pooled using wide bore tips, and nuclei transferred to a flow cytometry tube through a 0.35 μm filter cap and DAPI added to a final concentration of 3 μM. Pooled nuclei were then sorted on a FACS Aria II cell sorter (BD) at 150 cells per well into 96 well LoBind plates containing 5 μL of EB buffer (Qiagen). After sorting, 0.75 μL of second strand mix (0.5 μL of mRNA second strand synthesis buffer and 0.25 μL of mRNA second strand synthesis enzyme, New England Biolabs) were added to each well, second strand synthesis performed at 16C for 150 minutes. Tagmentation was performed by addition of 5.75 μL of tagmentation mix (0.01 μL of a custom TDE1 enzyme in 5.74μL 2x Nextera TD buffer, Illumina) and plates incubated for 5 minutes at 55C. Reaction was terminated by addition of 12 μL of DNA binding buffer (Zymo) and incubated for 5 minutes at room temperature. 36 μL of Ampure XP beads were added to every well, DNA purified using the standard Ampure XP protocol (Beckman Coulter) eluting with 17 μL of EB buffer and DNA transferred to a new 96 well LoBind plate. For PCR, 2 μL of indexed P5, 2 μL of indexed P7 (*63*) and 20 μL of NEBNext High-Fidelity master mix (New England Biolabs) were added to each well and PCR performed as follows: 75C for 3 minutes, 98C for 30 seconds and 18 cycles of 98C for 10 seconds, 66C for 30 seconds and 72C for 1 minute followed by a final extension at 72C for 5 minutes. After PCR, all wells were pooled, concentrated using a DNA clean and concentrator kit (Zymo) and purified via a 0.8X Ampure XP cleanup. Final library concentrations were determined by Qubit (Invitrogen), libraries visualized using a TapeStation D1000 DNA Screen tape (Agilent) and libraries sequenced on a Nextseq 500 (Illumina) using a high output 75 cycle kit (Read 1: 18 cycles, Read 2: 52 cycles, Index 1: 10 cycles and Index 2: 10 cycles).

2.6.10 Preparation of sci-RNA-seq3 libraries

Frozen nuclei were thawed as before and three-level sci-RNA-seq libraries prepared as described in Cao et al. (*59*). Nuclei were pelleted at 500 x g for 5 minutes, washed three times with NSB

and a small aliquot of nuclei stained with 0.4% trypan blue (Sigma-Aldrich) and nuclei counted using a hemocytometer. 80000 nuclei in 22 μL of NSB, 2 μL of 10 mM dNTP mix and were then distributed into a skirted 2 μL of ligation compatible indexed oligo-dT primers were distributed into each well of 96 well LoBind plates, incubated at 55C for 5 minutes and placed on ice. 14 μL of reverse transcription mix (8μL of Superscript IV first-strand buffer, 2 μL of 100 mM DTT, 2 μL of Superscript IV and 2 μL of RNAseOUT recombinant ribonuclease inhibitor) was then added to every well and RT performed on a thermocycler using the following program: 4C for 2 minutes, 10C for 2 minutes, 20C for 2 minutes, 30C for 2 minutes, 40C for 2 minutes, 50 for 2 minutes and 55C for 15 minutes. After RT, 60 μL of nuclei buffer containing BSA (NBB, 10 mM Tris-HCl, pH 7.4, 10 mM NaCl, 3 mM MgCl2 and 1% BSA) were added to each well, nuclei pooled using a wide bore tip, nuclei pelleted by centrifugation at 500 x g for 10 minutes and the supernatant removed. A second round of combinatorial indexing was performed by ligation of indexed primers onto the 5' end of RT indexed cDNA. Nuclei were resuspended in NSB and 10 μL added to each well of 96 well LoBind plates after which 8 μL of indexed ligation primers were added to each well along with 22 μL of ligation mix (20 μL of Quick ligase buffer and 2 μL of Quick ligase, New England Biolabs). Ligation was then performed at 25C for 10 minutes. After ligation, 60 μL of NBB were added to each well, nuclei pooled using a wide bore tip, another 40 mL of NBB added to the nuclei and nuclei pelleted by centrifugation at 600 x g for 10 minutes and the supernatant removed. Nuclei were then washed once with 5 mL of NBB, resuspended in 4 mL of NBB, multiplets removed by filtering using a 40 μm Flowmi cell strainer (Sigma-Aldrich), nuclei counted and 5000 nuclei were distributed per well into 96 well LoBind plates in a 5 μL volume. Plates containing nuclei were frozen and stored at -80C until further processing. After thawing the frozen plate 5 μL of second strand synthesis mix (3 μL of elution buffer, 1.33 μL mRNA second strand synthesis buffer and 0.66 μL of mRNA second strand synthesis enzyme) were added to each well and incubated at 16C for 3 hours. Tagmentation was performed by addition of 10 μL of tagmentation mix (0.01 μL of a custom TDE1 enzyme in 9.99μL of 2x Nextera TD buffer, Illumina) and plates incubated for 5 minutes at 55C. After tagmentation, 20 μL of DNA binding buffer was added to every well and plates incubated at room temperature for 5 minutes. 40 μL of Ampure XP beads were then

added to each well and plates incubated for 5 minutes at room temperature. Upon isolation of beads using a magnetic stand, supernatant was removed and beads were washed twice with 80% ethanol. 10 μL of USER reaction mix (1 μL of 10X USER buffer and 1 μL of USeR enzyme in nuclease-free water, New England Biolabs) was then added to each well and beads resuspended and incubated at 37C for 15 minutes. After incubation, 7 μL of elution buffer were added to each well and supernatant transferred to a new 96 well LoBind plate after binding beads on a magnetic stand. After incubation at 85C for 10 minutes, libraries were generated with 15 cycles of PCR. Following PCR amplification, sequencing library was purified by first concentrating 1mL of PCR library using a 1x Ampure cleanup and then running the resulting product on a 2% agarose gel containing ethidium bromide. Gel was cut to isolate 2 fragments, hash molecules (220bp - 250bp) and RNA library (250bp - 1000bp). Following gel extraction and an additional 1x Ampure cleanup RNA libraries were sequenced on a NovaSeq 6000 (Illumina) (Read 1: 34 bp, Read 2: 100 bp, Index 1: 10 bp and Index 2: 10 bp) and hash libraries were sequenced on a 75 cycle NextSeq (Read 1: 34 bp, Read 2: 38 bp, Index 1: 10 bp and Index 2: 10 bp).

2.6.11 Preparation of bulk RNA sequencing libraries

Compound treated cells were first trypsinized and harvested as described previously. Cells were then lysed in V-bottom plates using 26 μL of NSB. 2 μL of 25 μM indexed RT primers were added and annealed at 65C for 5 minutes. Subsequently, RT reaction was performed using the SuperScript IV system, with 8μL of 5x SuperScript Buffer, 2μL of SuperScript IV, 2μL 10 mM dNTP mix, 2μL of 100 mM DTT and 2 μL of RNAseOUT recombinant ribonuclease inhibitor per well.. Reaction was performed for 10 minutes at 55C and subsequently stopped via heat inactivation (80C for 10 minutes). Libraries were then pooled and excess RT primer was removed through either two 0.7x SPRI clean-ups or a single 0.7x SPRI cleanup followed by Exo-1 treatment and inactivation. Double stranded DNA was produced through incubation at 16C for 3 hours with second strand synthesis mix containing 0.5 μL of enzyme and 2 μL of second strand reaction buffer in a final volume of 20 μL. Following second strand synthesis, libraries were tagmented

with 1 μL of commercial Nextera reagent with 20.5 μL of 2x TD buffer and . Reactions were stopped with 40 μL of Zymo Clean and Concentrate buffer and incubated at room temperature for 5 minutes. Libraries were subsequently purified with a 1x SPRI cleanup and eluted in 16 μL of elution buffer. Sequencing libraries were generated through PCR with 2 μL of index P7 and P5 primers each and 20 μL of 2x NEB Next Master Mix. Finally, libraries were pooled, purified with a 1x SPRI cleanup and quantified. Libraries were sequenced on a Nextseq 500 (Illumina) using a high output 75 cycle kit (Read 1: 18 cycles, Read 2: 52 cycles, Index 1: 10 cycles and Index 2: 10 cycles).

2.6.12 Pre-processing of sequencing data

Sequencing runs were first demultiplexed using bcl2fastq v.2.18. Only barcodes that matched reverse transcription indices within an edit distance of 2 bp were retained. For sci-RNA-seq3 libraries, barcodes which matched both provided reverse transcription indices and ligation indices within an edit distance of 2 bp were retained. Following assignment of indices, polyA tails were trimmed using trim-galore, and reads were mapped to a human transcriptome (hg-38) or human-mouse transcriptome (hg-38 and mm-10) using the STAR aligner. Following alignment, reads were filtered for alignment quality, and duplicates were removed. Reads were considered duplicates if they (1) mapped to the same gene, (2) mapped to the same cell barcode and (3) contained the same unique molecular identifier (UMI). Reads that met the first two criteria, and differed by an edit distance of 1 from a previously observed UMI were also marked as duplicates and discarded. Non-duplicate reads were assigned to genes using bedtools (*78*) to intersect with an annotated gene model. All 3' UTRs in the gene model were extended by 100 bp to account for the possibility that some gene 3' UTR annotations may be too short, causing genic reads to improperly be annotated as intergenic. Cell barcodes were considered to correspond to a bona fide cell if the number of unique reads associated with the barcode was greater than an interactively defined threshold on a knee plot. Reads from cells that passed this UMI count threshold were first aggregated into a sparse matrix format and then loaded and saved as a CDS object for analysis with Monocle3.

2.6.13 Assigning sample labels from hash reads

Demultiplexed reads that matched combinatorial indexing barcodes were examined to identify hash reads. Reads were considered hash reads when they met two criteria: (1) the first 10 bp of read 2 matched a hash barcode in the experiment within an edit distance of two and (2) contained a polyA track between base pairs 12 to 16 of read 2. These reads were then deduplicated by cell barcode and collapsed by UMIs to create a vector of hash oligo UMI counts for each nucleus in the experiment.

To assign each nucleus to the culture well from which it came, we test whether its sci-RNA-seq library is enriched for a particular hash barcode. We compare a nucleus's hash UMIs against a 'background distribution', which under ideal circumstances, would be the uniform distribution. In practice, minor variation in concentrations of hash oligos added to each well of liberated nuclei may necessitate empirically estimating the background. To do so, we simply average the relative hash UMIs from cell indices for which fewer than $<$ mRNA UMIs were collected, reasoning that these reflect library contributions from RT well supernatant, debris fragments, etc. We then compare the hash UMIs for nucleus to this background by a chi-squared test. After correcting the resulting p values for multiple testing by Benjamini-Hochberg, we reject the null hypothesis that originates from the background distribution at specified FDR (5% FDR was used in this study). Those nuclei with hash counts deemed different than background are then evaluated for enrichment for a single hash sequence. Enrichment ratios were calculated as the UMI count ratio of the most abundant vs. the second most abundant hash oligo. Specifically, if the UMI count for the most abundant hash in nucleus is α-fold higher than the second most abundant, is marked as a singleton. α was determined on a per-experiment basis by examining the distribution of these ratios and choosing a value that separated unlabeled cells and singularly labeled cells. Cells that fell below α-fold enrichment of a unique hash oligo were flagged as a multiplet or debris and discarded.

2.6.14 Dose-response analysis

Dose-response analysis was conducted in R using the drc package (*79*) by fitting a four-parameter log-logistic model for each drug to the number of cells recovered in the single-cell RNA-seq data at each dose. Cells that survived doublet analysis and QC were grouped by their culture well of origin and counted. These counts were then adjusted to account for variation in recovery as a function of cell type and culture plate as follows. The vector of cell counts across wells were fit with the model

The dose response curves enable cells to be annotated according to the impact of their culture conditions on viability. Each cell is assigned a 'viability score' which is simply the expected fraction of vehicle cells remaining after exposure to a given dose of a compound. These cell counts are generated via the predict() function of the drc package and then normalized relative to the corresponding vehicle control.

2.6.15 Dimensionality reduction and trajectory analysis

Gene expression profiles were visualized with Monocle 3, which uses UMAP to project them into a two or three dimensional space. Briefly, Monocle 3 first calculates size factors for every cell. Size factors were calculated as the log UMI counts observed in a single cell divided by the geometric mean of log UMI counts from all measured cells. After scaling each nucleus' UMI counts by its library size factor, Monocle3 adds a pseudocount of 1, and log transforms the counts. Next, these log-transformed profiles are projected onto the top 25 principal components. These PCA coordinates were transformed by Monocle 3 (using an approach similar to the removeBatchEffect() function in the limma package (*80*)) according to the model l̃og(UMIs) + replicate (Figure 3) or ' log(UMIs) + viability + proliferation index + replicate' . Adjusted PCA coordinates for each cell are used to initialize UMAP. Unless otherwise noted, UMAP was run with the following parameters: 50 nearest neighbors, min_dist = 0.1, inter-cell distance assessed by cosine similarity. UMAP projection of cells after dual HDAC inhibition and acetyl-COA precursor supplementation or acetyl-CoA generating enzyme inhibition was performed as described with the exception that

PCA initialization was performed on the top 1000 most overdispersed genes. Louvain community detection was then performed on this UMAP space using the python package 'louvain'. Trajectory reconstruction was then performed as described in (*59*).

To determine whether cells exposed to a particular compound/dose combination displayed an enrichment along UMAP space we created contingency tables of the number of compound or vehicle treated cells within and outside clusters and used the stats R package implementation of Fisher's exact test to test for enrichment. For visualization of drug enrichment in Figure 3B, cells opacity was added to cells under the minimum compound/dose that passed meet an enrichment cutoff of FDR $<$ 1% and a log2 of the odds ratio ¿ 2.5. Cells that passed these filters were used to generate the heatmap of the fraction of enriched cells by cluster in Figure S6.

2.6.16 Estimation of Proliferation Index

To obtain an estimate of proliferation index for a single cell, size factor normalized expression of cell cycle marker genes (from Table S5 in Tirosh et. al. Science, 2016) were summed for each cell and logged. Scores were calculated in this way for both G1S and G2M. "Proliferation Index" refers to overall proliferative state of a cell and is calculated as the logged sum of the aggregated G1S and G2M gene expression.

2.6.17 Differential expression analysis

To test whether a gene is differentially expressed by a cell line in a dose-dependent manner when exposed to a compound, we fit its (library size-factor adjusted) UMI count recorded from each nucleus with a generalized linear model:

Where is a quasipoisson-valued random variable, is the log-transformed dose of the compound being evaluated. We fit these models with Monocle 3, which uses the speedglm package. To fit the regression model for each drug's effect on each gene, we first identify the subset of cells that are relevant for the model. To determine the effects on gene G in cells of type C when treated with drug D, we include all cells of type C that were treated with any dose of D. To these, we

add cells of type C that were treated with the vehicle control. We then fit a model defined above relating the expression level of G across all of these cells. Genes are deemed to be dose-dependent differentially expressed genes (DEGs) if their fitted models include a term that is significantly different from zero as assessed by a Wald test (Benjamini-Hochberg adjusted). P values for terms are pooled across all compounds and all genes prior to correction for multiple testing.

To assess a gene for differential expression as a function of 'pseudodose' in the consensus HDAC inhibition trajectory, we fit a model

Where is a quasipoisson variable capturing the gene's UMI counts, encodes the pseudodose values smoothed via a natural spline, is a factor encoding the cell type, and captures the interaction between cell type and pseudodose. The term encodes the (log) dose dependent effects of compound

.

2.6.18 Pairwise correlation of screened compounds

To identify compounds that result in similar dose-dependent changes to cellular transcriptomes we calculated the Pearson correlation between every pairwise set of compounds. We created a gene by compound matrix for the union of dose-dependent genes across all compounds where each entry is the beta coefficient for the dose dependence term and ten calculated the Pearson correlation for every drug pair using the cor.test() function in the R stats package specifying to use complete observations. The resulting correlation matrix was then hierarchically clustered using the pheatmap package in R. The significance of every pairwise correlation was determined using the corr.test() function from the psych package in R specifying Benjamini-Hochberg as the method for adjusting for multiple hypothesis testing.

2.6.19 Geneset enrichment analysis

After fitting a generalized linear model, genes that had significant coefficients (5% FDR threshold) were used for gene set enrichment analysis with the R package piano (*81*). Briefly, gene sets were ranked according to the set-wide average Wald test statistic corresponding to the generalized linear

model term being evaluated with piano's runGSA() function. Genes were randomized across sets to establish a null distribution for each set's rank. After 10000 permutations, runGSA() computed p values using the 'mixed' directional enrichment policy.. The top gene sets, corresponding to those with the largest magnitude enrichment statistic, were chosen for visualization.

2.6.20 Alignment of HDAC inhibitior treated cells

To organize cells treated with HDAC inhibitors into a trajectory cells were sampled to equalize the number of cells represented between the three cell lines or between treatments at 24 and 72 hrs. Next, PCA coordinates were computed jointly, and then aligned using the mnnCorrect function from the package scran (*70*). These adjusted coordinates were used to initialize UMAP in Monocle 3. We then fit a principal graph to the data via lean_graph(). To define the origin of the trajectory, we mapped each cell to its nearest principal graph node, and then selected all principal graph nodes for which a majority of mapped cells were treated with vehicle. All other cells' pseudodoses was measured as the geodesic distance between their nearest principal graph node to an origin node.

To quantify the potency of each HDAC inhibitor, we first grouped all cells from each replicate according to treatment and dose, and then computed the mean pseudodose for each cell. We then fit mean pseudodose values as a function of compound concentration using the drc package (*79*). We used a four-parameter log-logistic model, with the maximal response fixed at the highest pseudodose value achieved across all compounds and doses. We then take the model parameter as described in the 'dose response analysis' section above as the transcriptional EC50 (TC50) for each compound.

2.7 Supplemental Figures

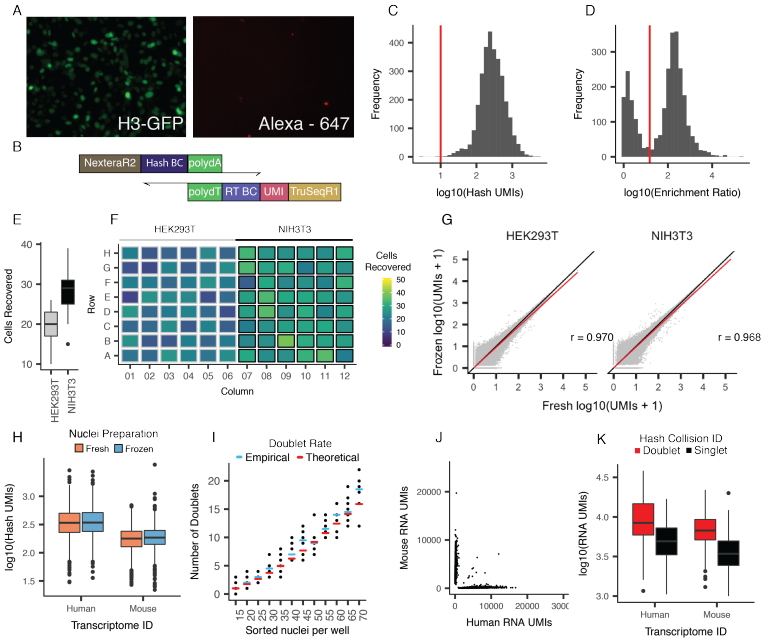

Figure 2.6: **Hashing with short, polyadenylated single-stranded oligonucleotides enables stable, low-cost labeling of nuclei for sci-RNA-seq and subsequent doublet detection (A)** Fluorescent microscopy images demonstrating lack of Alexa 647-conjugated oligo staining (right) of unpermabilized H3-GFP NIH3T3 cells (left). **(B)** Design of polyadenylated hash oligos (top) and indexed primer used for reverse transcription (bottom). **(C)** Number of hash UMIs detected per cell. Cells with fewer than 10 hash UMIs (red line) were excluded from further analysis. **(D)** Distribution of enrichment ratios for cells. Enrichment ratios were calculated as the UMI count ratio of the most abundant vs. the second most abundant hash oligo. An enrichment ratio cutoff of 15 (red line) was used to distinguish doublets vs. singlets. **(E)** Boxplot of the number of cells recovered per well for each cell line. **(F)** Layout of culture plate wells with color indicating number of cells recovered and outline indicating cell line. Note that although more NIH3T3 cells were recovered per well, similar numbers of cells were recovered across wells of each cell type. **(G)** Log-scale per-gene aggregated, size-factor normalized UMI counts recovered from sci-RNA-seq on fresh vs. frozen preparations. Size factors are calculated as the log counts observed in a single cell divided by the geometric mean of log counts from all measured cells. Black line indicates y = x. Red line is the fit with Pearson correlation shown. **(H)** Log-scale boxplot of number of hash UMIs recovered from sci-RNA-seq of HEK293T (human) or NIH3T3 (mouse cells) from fresh vs. frozen preparations. **(I)** Theoretical (red bars) vs. observed (black dots for individual wells and blue bars for means) doublet rate as a function of the number of nuclei sorted into the final plate during sci-RNA-seq. **(J)** Barnyard plot from Figure 1E after removal of doublets detected by hashing. **(K)** Log-scale boxplot of number of RNA UMIs in singlet vs. doublet cells, as called based on the purity of hash UMIs. Of note, these are 'within species' doublets, i.e. human-human or mouse-mouse, which are not readily detected by conventional barnyard experiments.

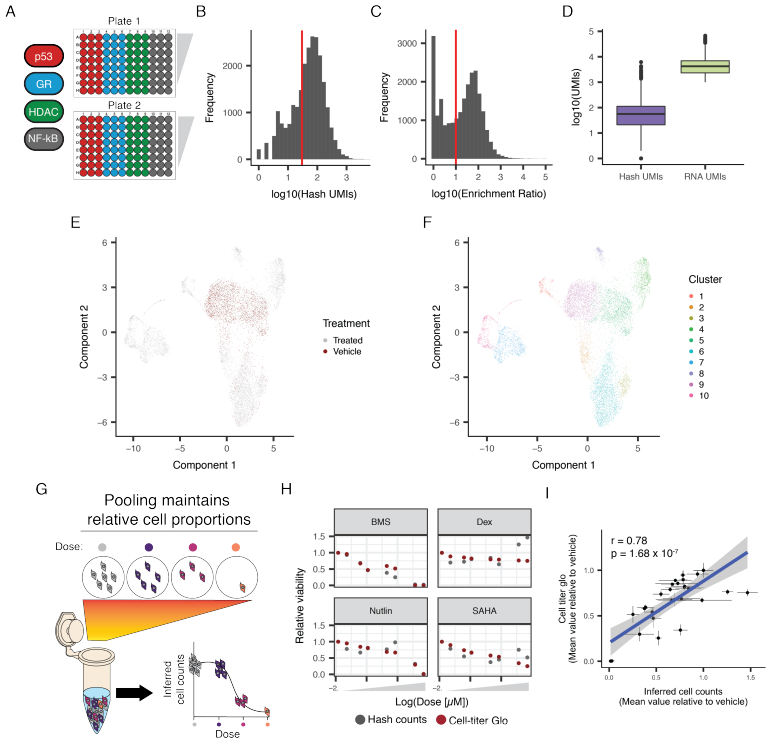

Figure 2.7: **sci-Plex distinguishes transcriptional responses of A549 cells to four small molecules and recovers dose-response estimates similar to established assays** A) Experimental layout of A549 cells in 96 well plates. Cells were treated for 24 hours in two 96 well plates using 7 doses (or vehicle) arrayed along each column. B) Cells that contained more than 30 hash oligo UMIs and C) had an enrichment ratio of greater than 10 were retained. D) Retained cells had a median hash UMI count of 78 and median RNA UMI count of 4,681. E) UMAP embedding of chemically perturbed A549 cells, equivalent to Figure 2B but with cells colored by whether they were treated with vehicle or one of the four small molecules. F) UMAP embedding of chemically perturbed A549 cells, equivalent to Figure 2B but with cells colored by cluster as defined using the density peak algorithm in Monocle 3. G) Cartoon depicting how pooling of barcoded nuclei preserves relative cell counts. H) Viability estimates from counting the proportion of recovered hashed nuclei (grey) vs. CellTiter-Glo (red, n = 6). I) Scatter plot of inferred cell counts (x-axis) and CellTiter-Glo viability estimates (y-axis) across all treatments and doses tested (Pearson correlation and chi square test).

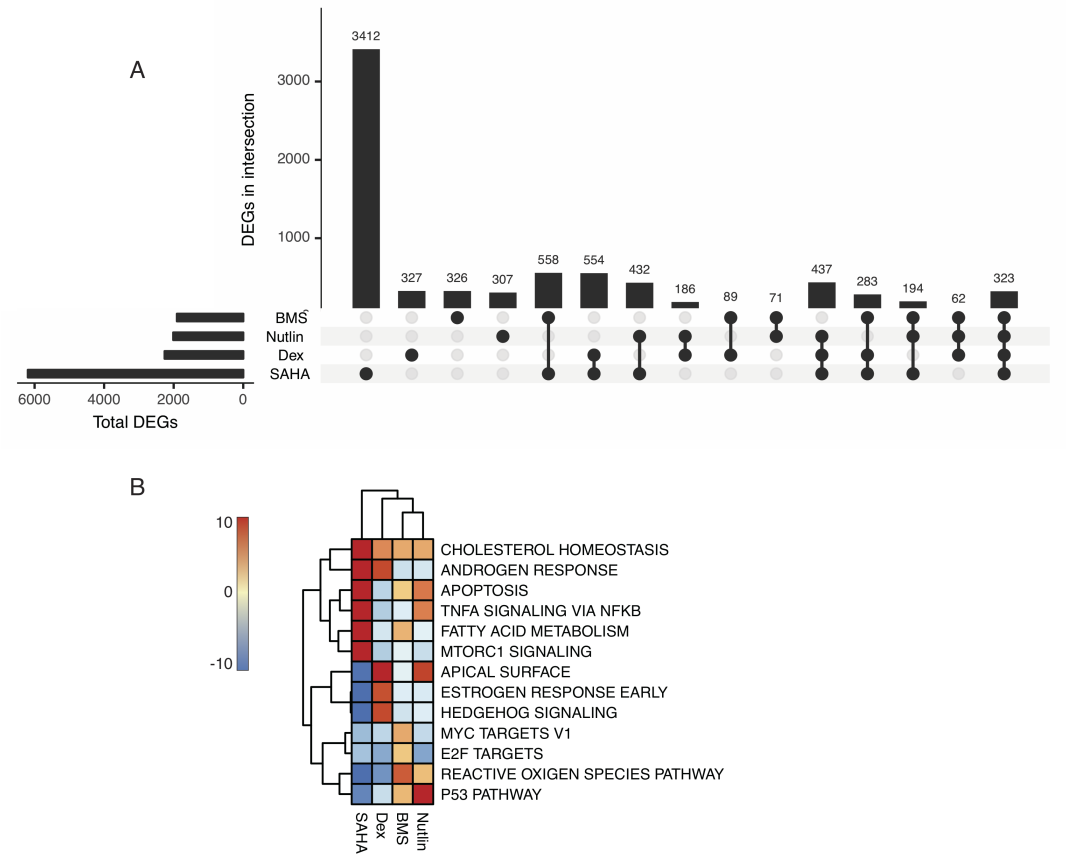

Figure 2.8: **Dose-dependent differentially expressed genes (DEG) recover expected transcriptional modules.** A) Upset plot displaying the intersections of dose-dependent DEGs between treatments (vertical bars) as well as the total number of dose-dependent DEGs per treatment (horizontal bars). A gene is defined as a dose-dependent DEG if the quasi-poisson regression model relating its expression in a given cell to the dose of drug that cell received shows a significant dose effect (Wald test) after Benjamini-Hochberg correction (FDR < 0.05). See Methods for full details on regression modeling. The four leftmost vertical bars correspond to drug-specific dose-dependent DEGs, while the rightmost vertical bar corresponds to dose-dependent DEGs shared by all four drugs. B) Gene set analysis (GSA) performed with dose-dependent DEGs using the runGSA() function from the piano package and the Hallmarks gene set from MSigDB (*82*). Heatmap color indicates the value of the directional GSA enrichment statistic with values that were capped at either -10 or +10 for visualization.

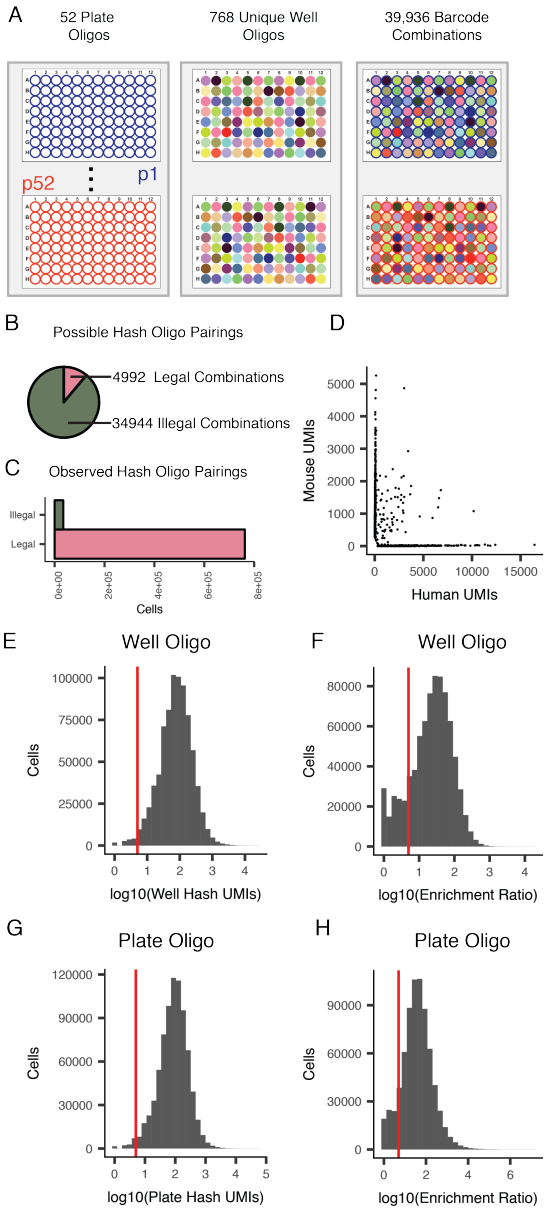

Figure 2.9: **Hash-based cell labeling in large-scale sci-Plex experiment.** A) Hashing design for sci-Plex with 188 compounds. The experiment used 52 x 96-well plates where each well was marked by a combination of two oligos, one specific to a single 96-well culture plate and another specific to a well within that culture plate. B) Although this could theoretically be implemented with just 96 well hash oligos, we instead used 768, which meant that out of the 39,936 possible pairings of plate and well hash oligos, only a minority (12.5%) of combinations were expected ('legal'), while most were unexpected ('illegal') C) Observed pairings of plate and well hash oligos were strongly enriched for 'legal' combinations. D) Scatter plot of HEK293T and NIH3T3 cells seeded in a single RT well of the large-scale sci-Plex experiment. E-H) Hash UMI (panels E,G) and enrichment ratio (panels F,H) cutoffs used for well hash oligos (panels E,F) and plate hash oligos (panels G,H). Enrichment ratio cutoffs corresponds to greater than 5-fold enrichment. Hash UMI cutoffs correspond to 5.

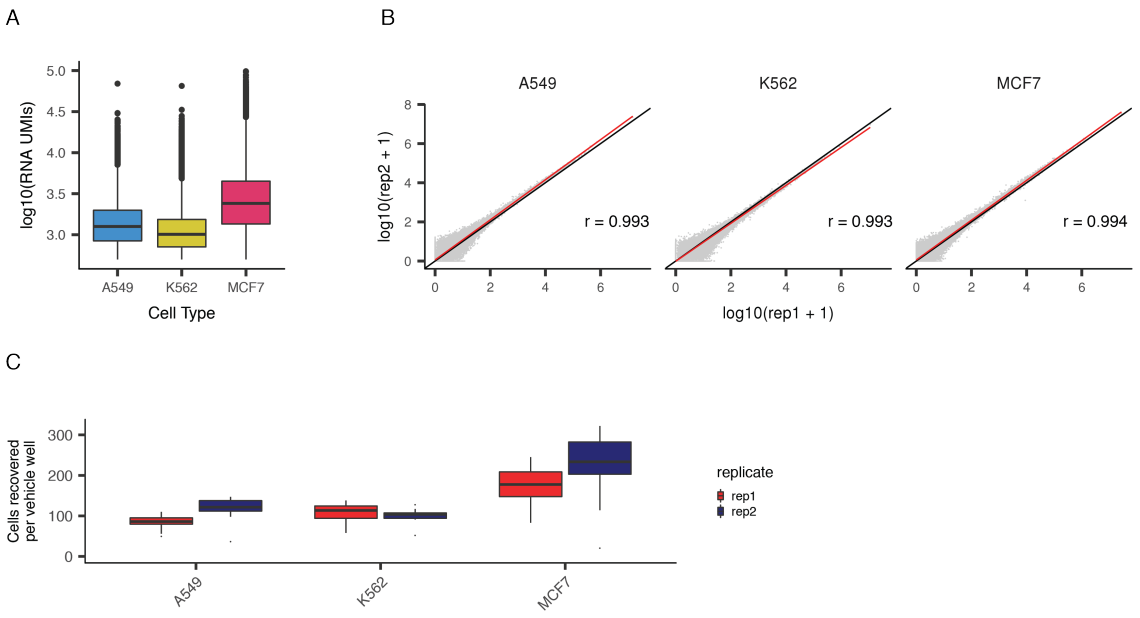

Figure 2.10: **Quality control metrics for large-scale sci-Plex experiment.** A) Log-scale boxplot of number of RNA UMIs for cells that passed hash and RNA UMI cutoff filters for each of three cell lines. B) Correlation of size factor-normalized counts for genes between replicates for each of the three cell lines. Black line indicates y = x. Red line is the fit with Pearson correlation shown. C) Boxplots showing the number of vehicle cells recovered from each of 8 vehicle control wells within each replicate for A549, K562 and MCF7 cells.

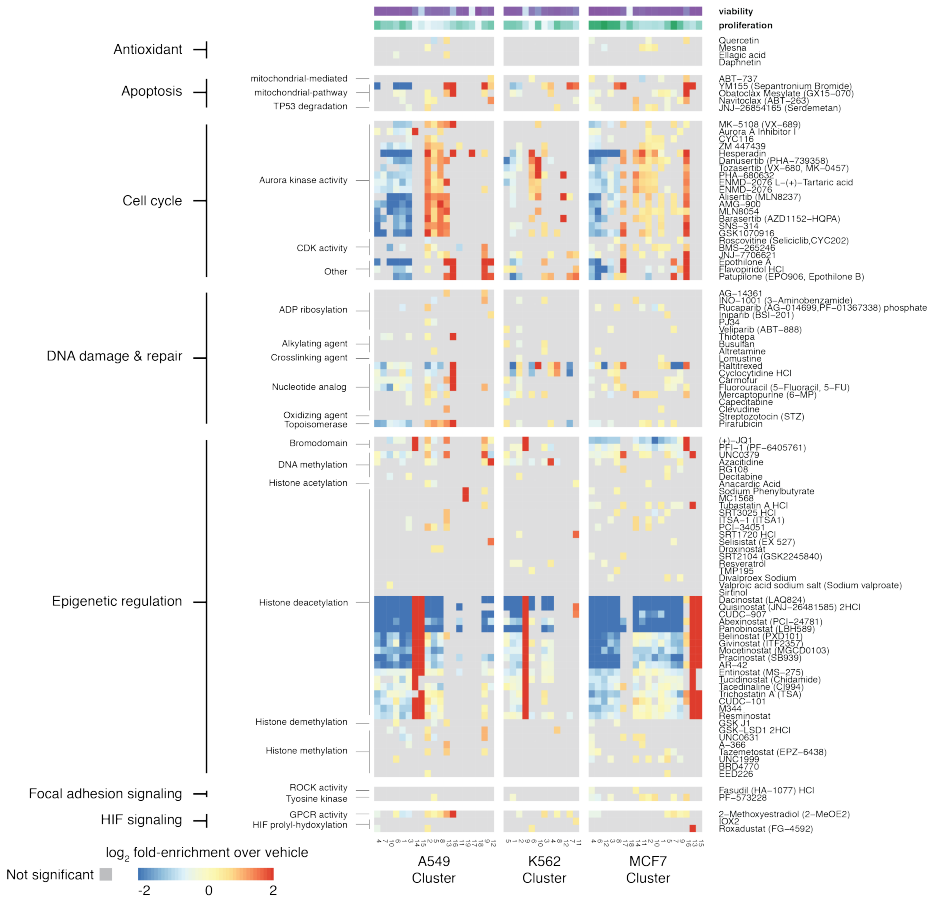

Figure 2.11: **Exposing cells to compounds alters their distribution across cell clusters.** Heatmap showing the log-transformed ratio of cells treated with a particular drug compared to vehicle control cells in each Louvain community. Columns correspond to clusters in PCA space (see **(Figure 3.12A-C**) and rows correspond to compounds, annotated by pathway and target. A gray entry denotes a compound that is not significantly enriched or depleted relative to vehicle in the corresponding cluster (Fisher's exact test, FDR $< 1\%$).

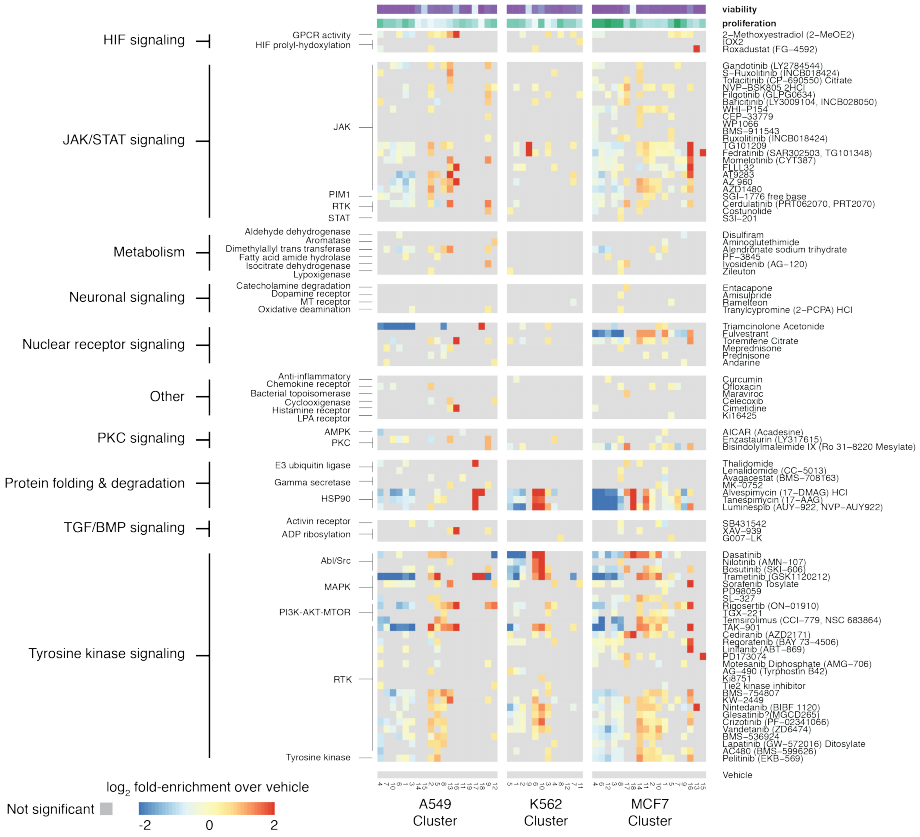

Figure 2.12: **Exposing cells to compounds alters their distribution across cell clusters.** Heatmap showing the log-transformed ratio of cells treated with a particular drug compared to vehicle control cells in each Louvain community. Columns correspond to clusters in PCA space (see **(Figure 3.12A-C**) and rows correspond to compounds, annotated by pathway and target. A gray entry denotes a compound that is not significantly enriched or depleted relative to vehicle in the corresponding cluster (Fisher's exact test, FDR $< 1\%$).

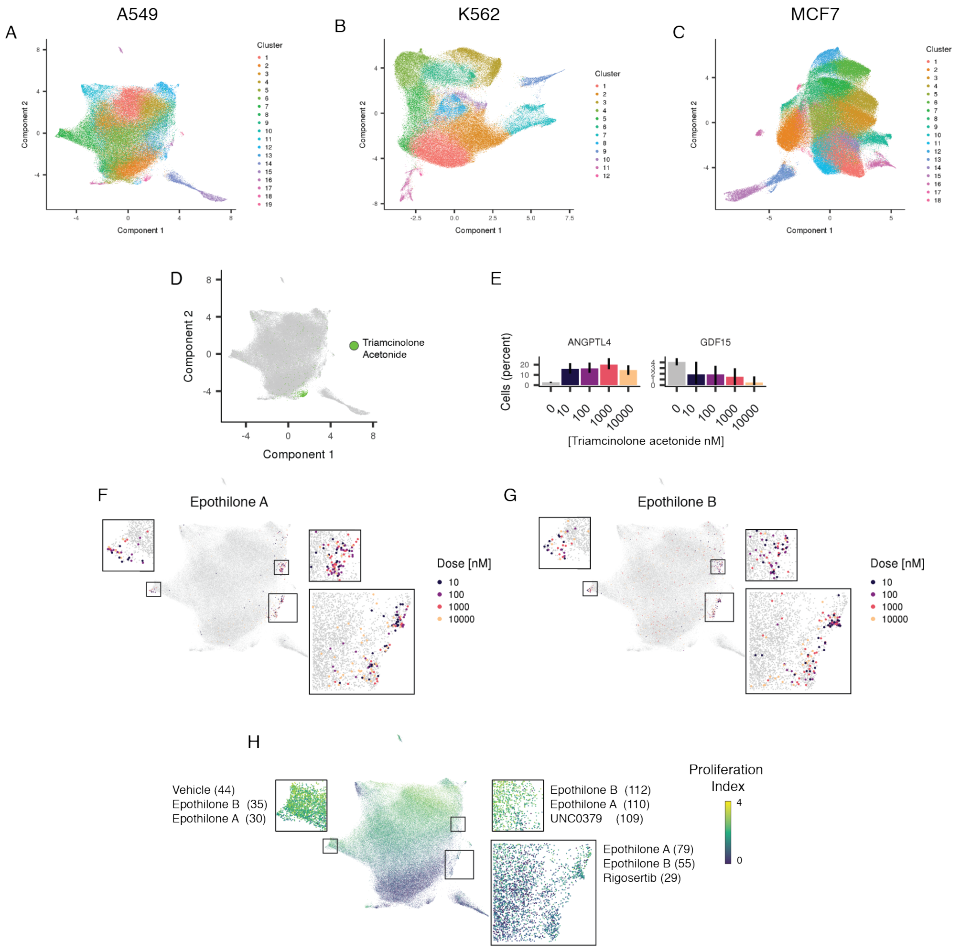

Figure 2.13: **sci-Plex identifies pathway-specific enrichment of compounds across UMAP clusters.** A-C) UMAP embedding from Figure 3B colored by cells' assignment to Louvain communities across PCA space for A549 (panel A), K562 (panel B) and MCF7 (panel C) cells. D) UMAP embedding of A549 cells from Figure 3B. Cells treated with the glucocorticoid receptor (GR) agonist triamcinolone acetonide are highlighted in green while all other cells are colored grey. These cells comprise the vast majority (95%) of the cells in cluster 18 from panel A. E) Percent of A549 cells expressing the GR target genes ANGPTL4 and GDF15, as a function of increasing doses of the synthetic GR agonist triamcinolone acetonide. F-H) UMAP embedding of A549 cells colored by cells treated with varying doses of Epothilone A (F), Epothilone B (G), or colored by proliferation index (H). Insets display magnified views of distinct foci induced upon treatment. The treatments with the highest number of cells in each bounding box are indicated in panel H with the number of cells in parentheses.

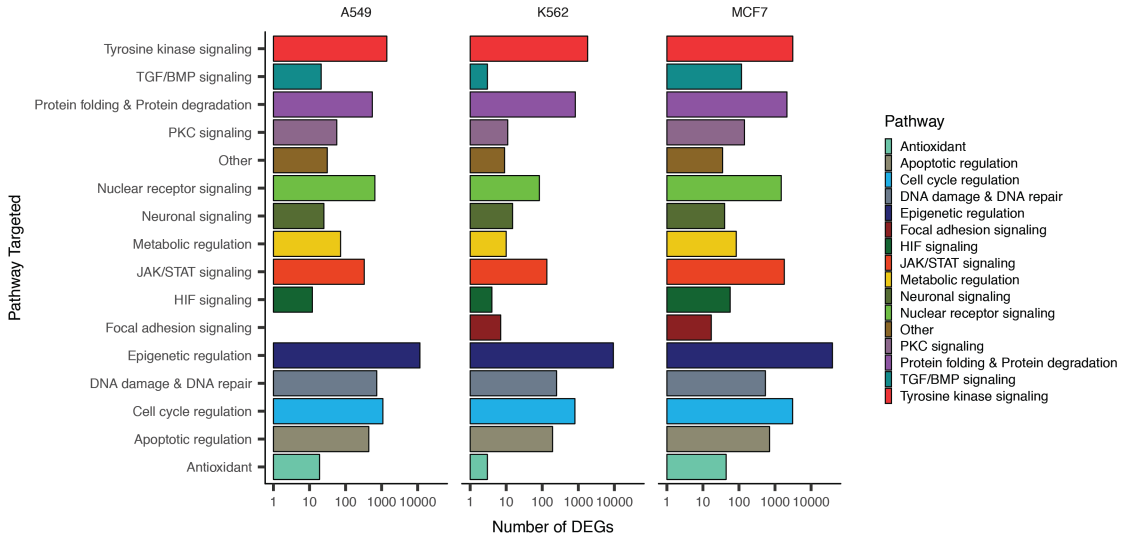

Figure 2.14: **Number of dose-dependent differentially expressed genes detected per compound category.** Significant dose-dependent differentially expressed genes (FDR < 0.05) are grouped by cell line and colored by targeted pathway.

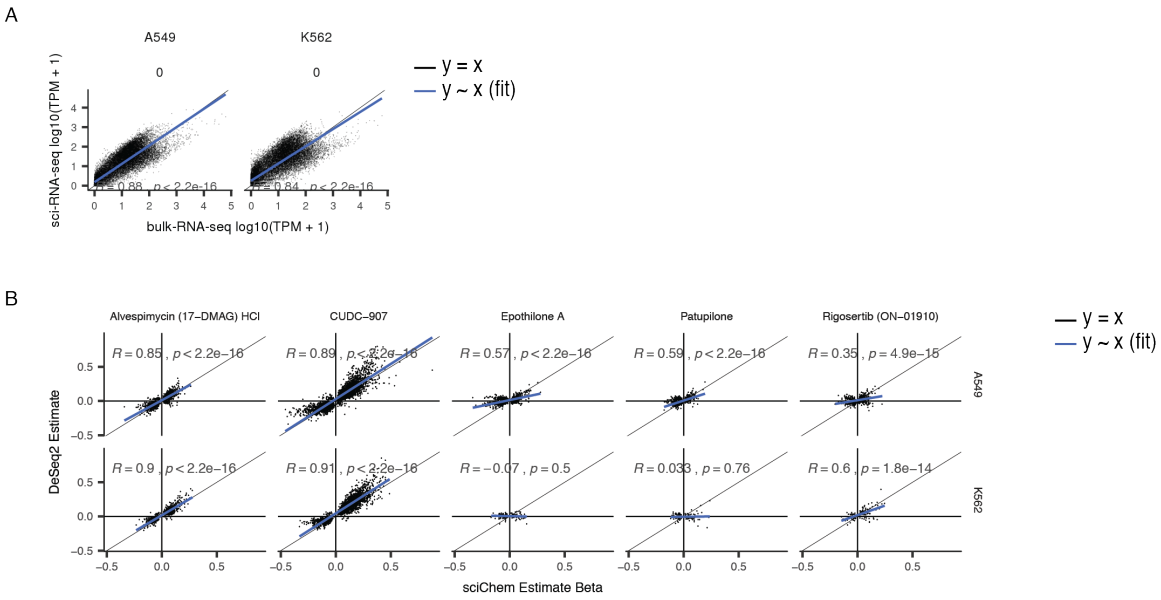

Figure 2.15: **Correlation of "pseudobulk" sci-Plex with bulk-RNA-seq.** A) Log10 transcripts per million (TPM) for protein-coding genes measured by bulk RNA-seq (x-axes) vs. size factor-normalized, aggregated single cell profiles for vehicle treated cells from sci-Plex (y-axis). Results are shown for both A549 and K562 cells. Black line indicates the line y = x, while the blue line shows the linear fit with Pearson correlation shown. B) Scatter plots, for selected compounds, comparing statistically significant estimates derived from linear models fit to single cell data (x-axes) vs. estimates derived from bulk RNA-seq using DESeq2 (y-axes). Black line indicates y = x. Blue line is the fit with Pearson correlation shown.

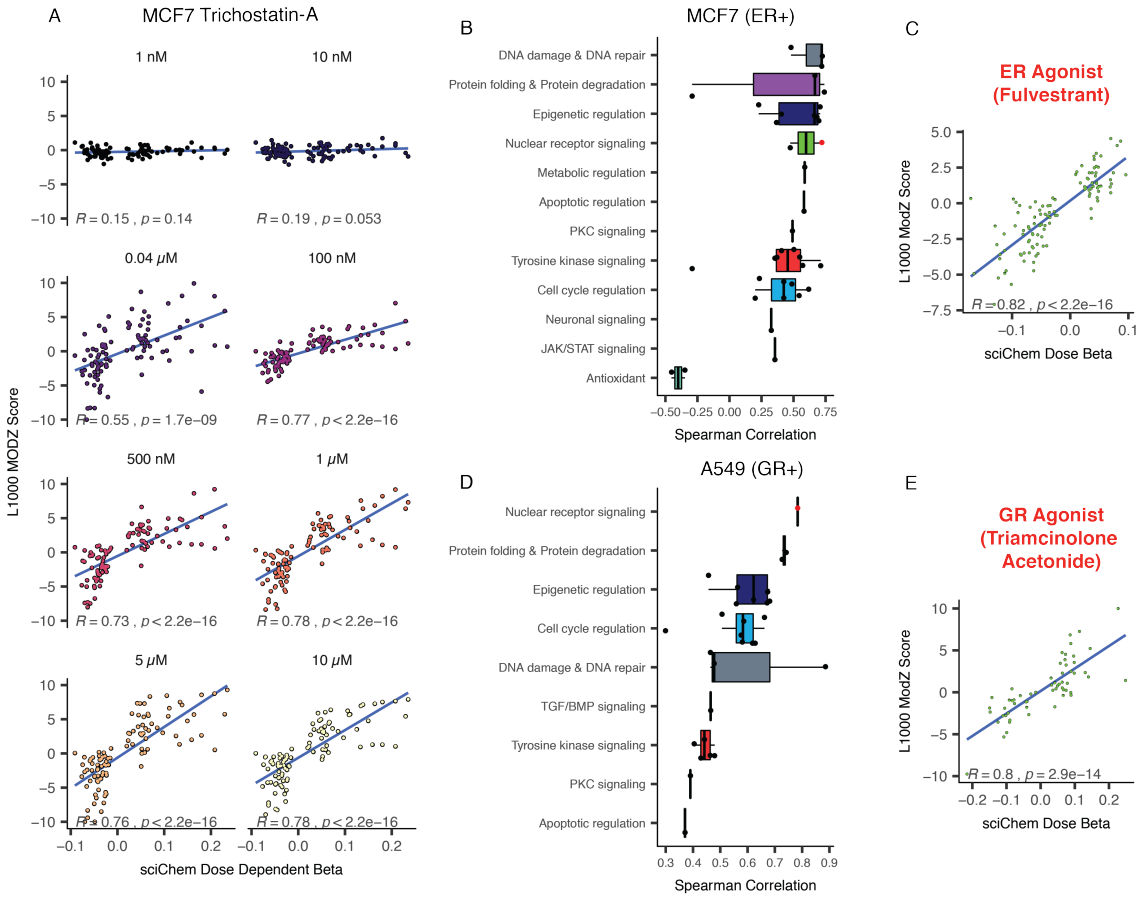

Figure 2.16: **Moderated Z scores from the L1000 assay correlate with dose-dependent betas from sci-Plex** A) For a selected compound-cell line combination (Trichostatin A in MCF7 cells), we plot moderated Z scores from the L1000 assay with treatment for 24 hrs at each of eight doses (y-axes) (*49*) vs. dose-dependent betas from sci-Plex data (x-axes). All genes that are part of the L1000 assay and significant for dose-dependent effects with sci-Plex (p-value < 0.01) are shown. Line is the fit with Spearman correlation shown. B) Boxplot of Spearman correlations between significant sci-Plex computed dose-dependent betas and L1000 moderated Z-score values from LINCS L1000 data for measured genes at the highest dose in MCF7 cells. Compounds are presented as grouped by the pathway they target. Red point corresponds to Fluvestrant. C) Similar to panel A, but for Fluvestrant in MCF7 cells and at the highest dose (10 μM). D) Similar to panel B, but for A549 cells. Red point corresponds to Triamcinalone Acetonide. E) Similar to panel A, but for Triamcinalone Acetonide in A549 cells and at the highest dose (10 μM).

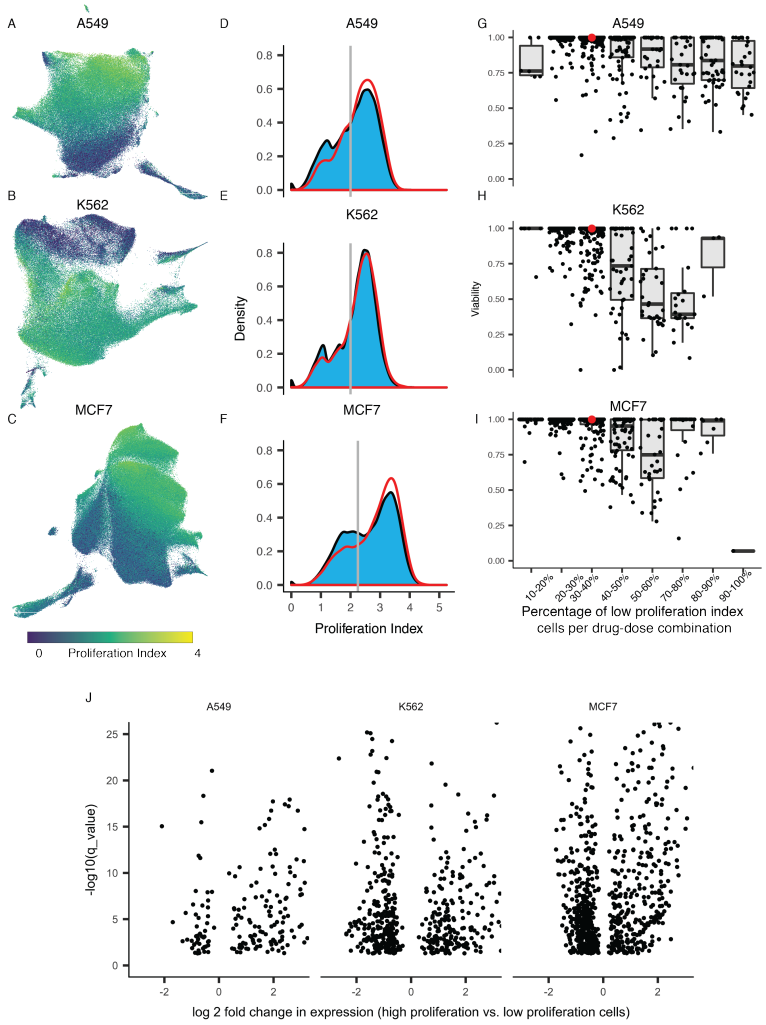

Figure 2.17: **Single cell measurements reveal variation in proliferation status in vehicle treated cell and across each dose of each drug.**A-C) UMAP projection of A549 (A), K562 (B) and MCF7 (C) colored by proliferation index. High proliferation index indicates an increase in the aggregate expression of transcripts that are markers for G1/S phase or G2/M phase (*83*). (D-F) Density plot of cell cycle distribution for compound-treated cells (blue fill) or vehicle-treated cells (red line). Grey line indicates cutoff used to distinguish proliferating cells (greater than cutoff) vs. non-proliferating cells (less than cutoff). G-I) Relationship between the percentage of cells designated as low proliferation at each dose of each drug (x-axis) versus the median estimated viability of that combination (y-axis). Each black point corresponds to cells treated with the same dose of a given drug. Red points correspond to vehicle treatment. J) Volcano plot depicting the log2 fold change for significant (q value < 0.01) differentially expressed genes between high and low fractions of vehicle treated cells.

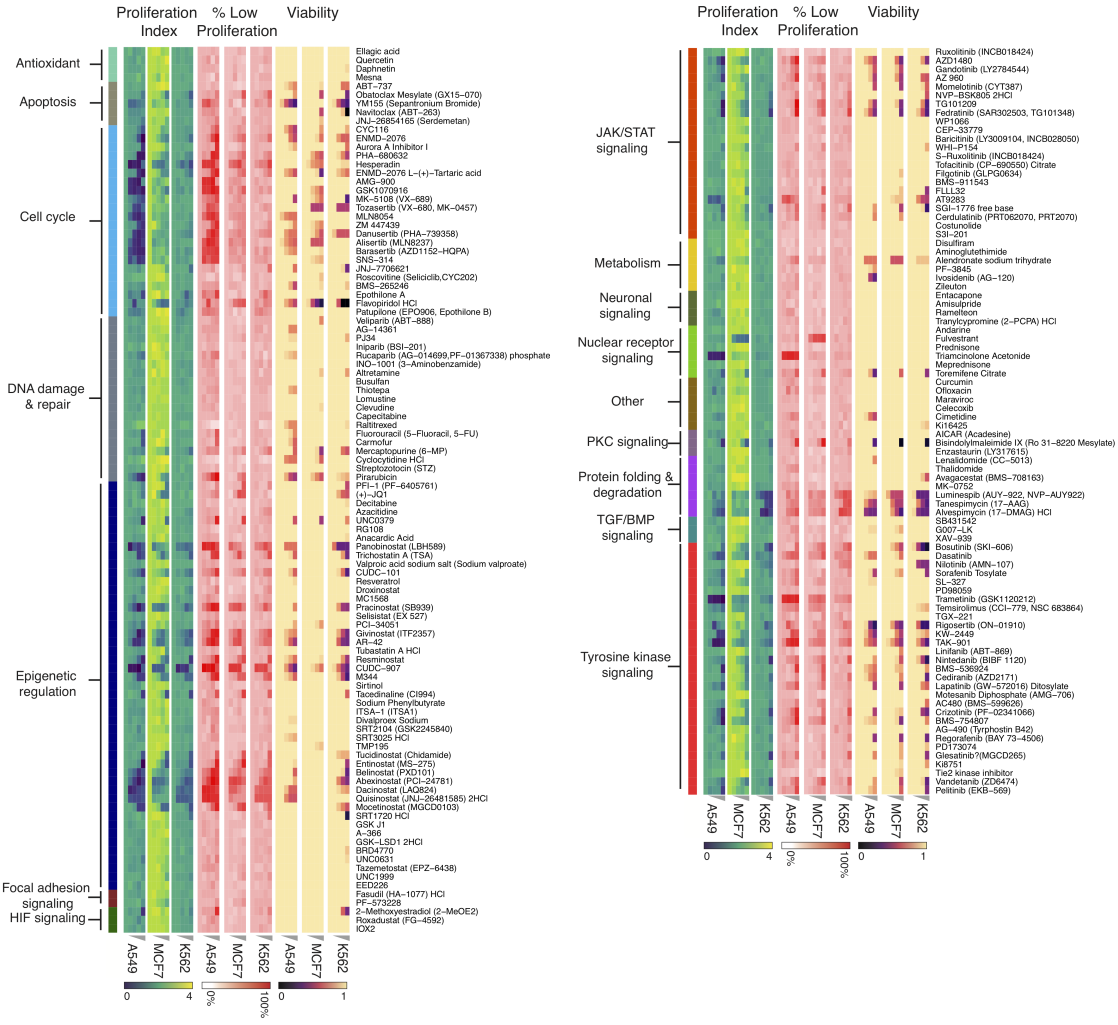

Figure 2.18: **Single cell measurements enable estimation of proliferation status and viability across drug-dose combinations.** Heatmap depicting estimates of relative proliferation rate, the percentage of cells exhibiting low proliferation index, and the estimated viability for each compound (row) at each dose (column) pair.

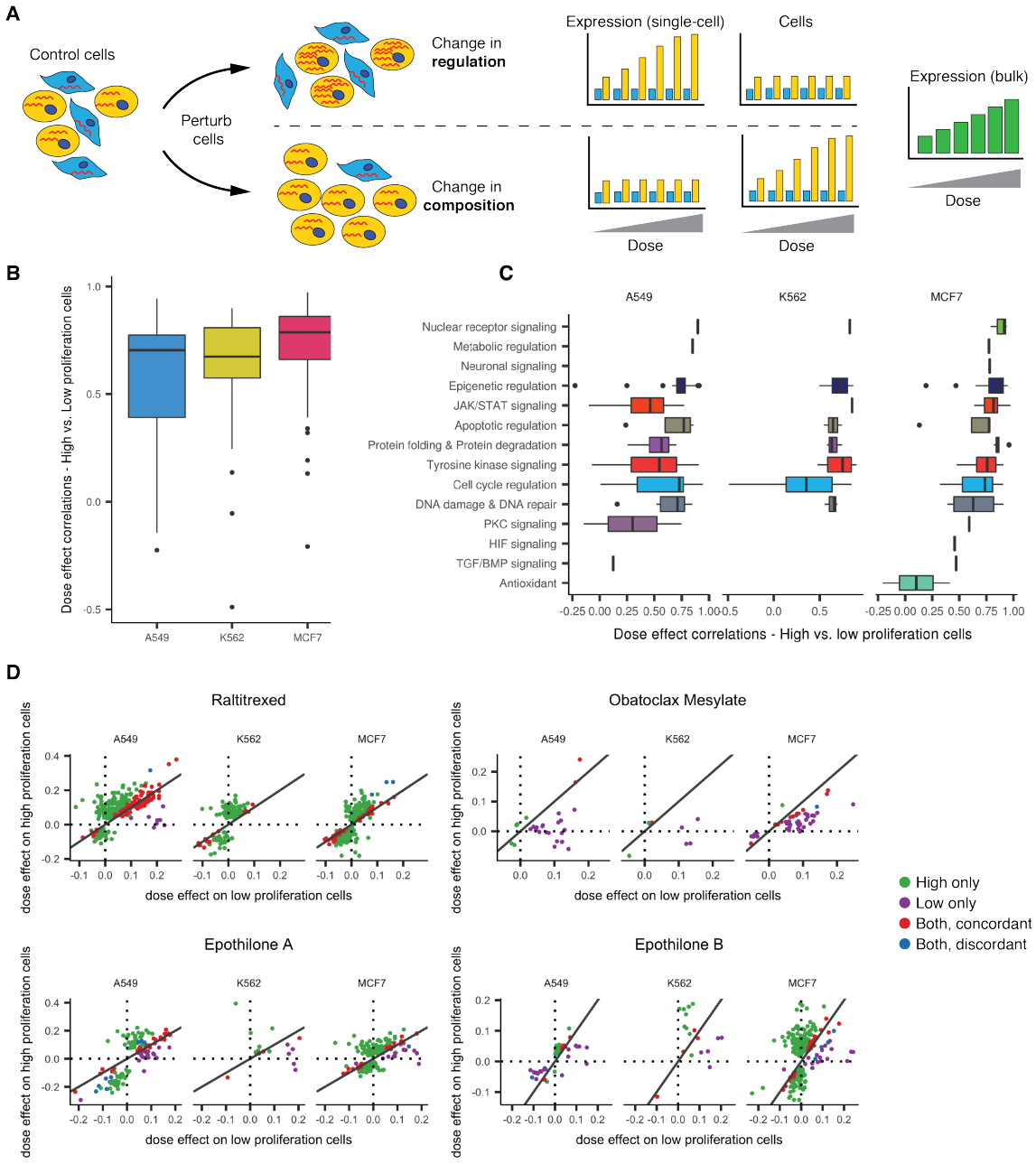

Figure 2.19: **sci-Plex enables the dissection of proliferating and non-proliferating cell populations.** A) Schematic depicting how changes in cellular state (top) and changes in the relative frequency of subpopulations (bottom) look identical upon subjecting the sample to aggregate measures such as bulk RNA-seq. Adapted from ref (*52*). B,C) Pearson correlations between dose-dependent effect sizes estimated from high vs. low proliferation index cells for each cell line (panel B) and drug class (panel C). D) Per-gene effect sizes estimated from high β_{dh} vs. low β_{dl} proliferation index cells for 4 selected compounds. Effect sizes are expressed as log2 transformed fold changes over intercept. Four classes of genes are shown: those significant in only high proliferation index cells (green); only low proliferation index cells (purple); both high and low cells, and with concordant effect estimates (red); both high and low cells, but with discordant effect estimates (blue). A drug had concordant dose-dependent effects on gene $\langle$ in high cells β_{dh} and low cells β_{dl} when $|\beta_{dh} - \beta_{dl}|$ was less than 10 percent of $\frac{1}{2}|\beta_{dh} - \beta_{dl}|$. Black line indicates $y = x$.

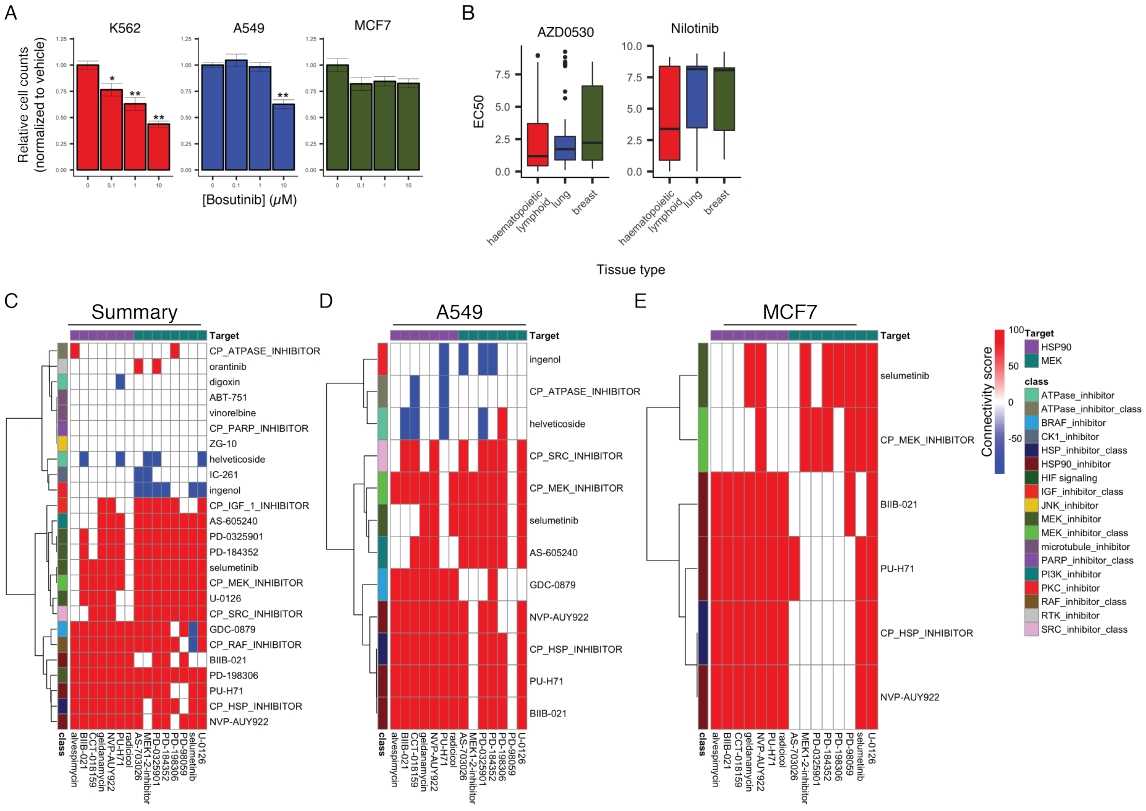

Figure 2.20: **sci-Plex screen identifies viability and expression signatures that are reproducible across validation experiments and orthogonal datasets.** A) Cell count viability estimates for K562 (red), A549 (blue) and MCF7 (green) cells exposed to vehicle or increasing doses of the Src/Abl inhibitor bosutinib (n = 6 culture replicates, Wilcoxon rank sum test). For each cell line, cell count values were normalized to the mean cell counts value of vehicle control treated cells. Error bars denote standard error of the mean, n = 8. B) EC50 values for cell lines of haematopoietic and lymphoid, lung and breast tissue origin, for which viability estimates are available from the Cancer Cell Line Encyclopedia (CCLE), exposed to the Abl inhibitors AZD0530 (left panel) or nilotinib (right panel). C-E) Top connectivity scores (a measure that summarizes similarities between transcriptional signatures induced by different drugs (*49*, *50*)) for MEK and HSP inhibitors from the CMAP database across all cell lines (C) or for A549 (D) and MCF7 (E) cells individually. A connectivity score cutoff of +/- 90 was applied as in (*49*)

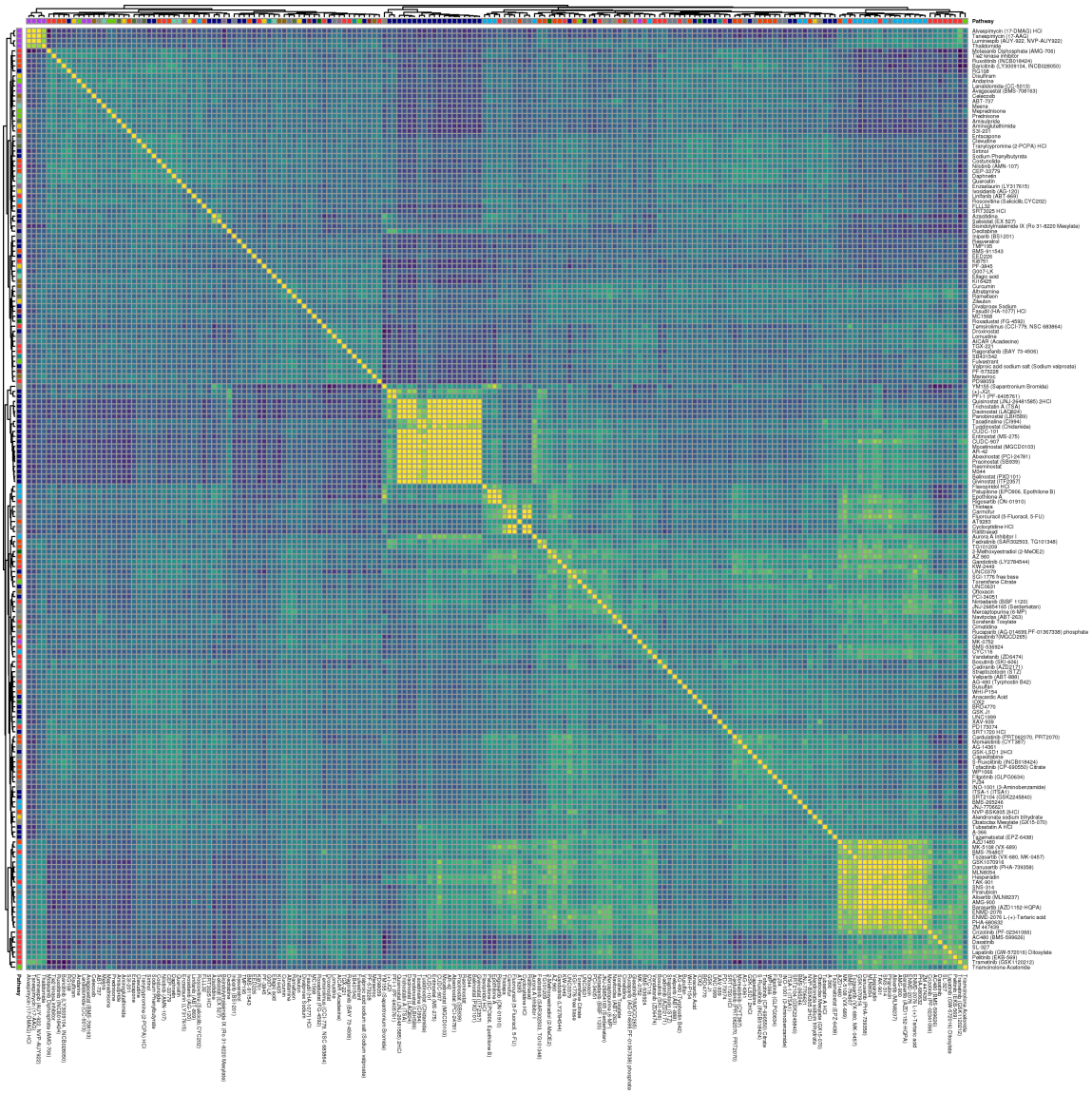

Figure 2.21: **Correlation of compound-driven molecular signatures for A549 cells identified in sci-Plex screen.** Heatmap depicts the Pearson correlation of beta coefficients across dose-dependent differentially expressed genes for every pairwise combination of compounds screened. To aid in visualization Pearson correlations were capped at 0.6.

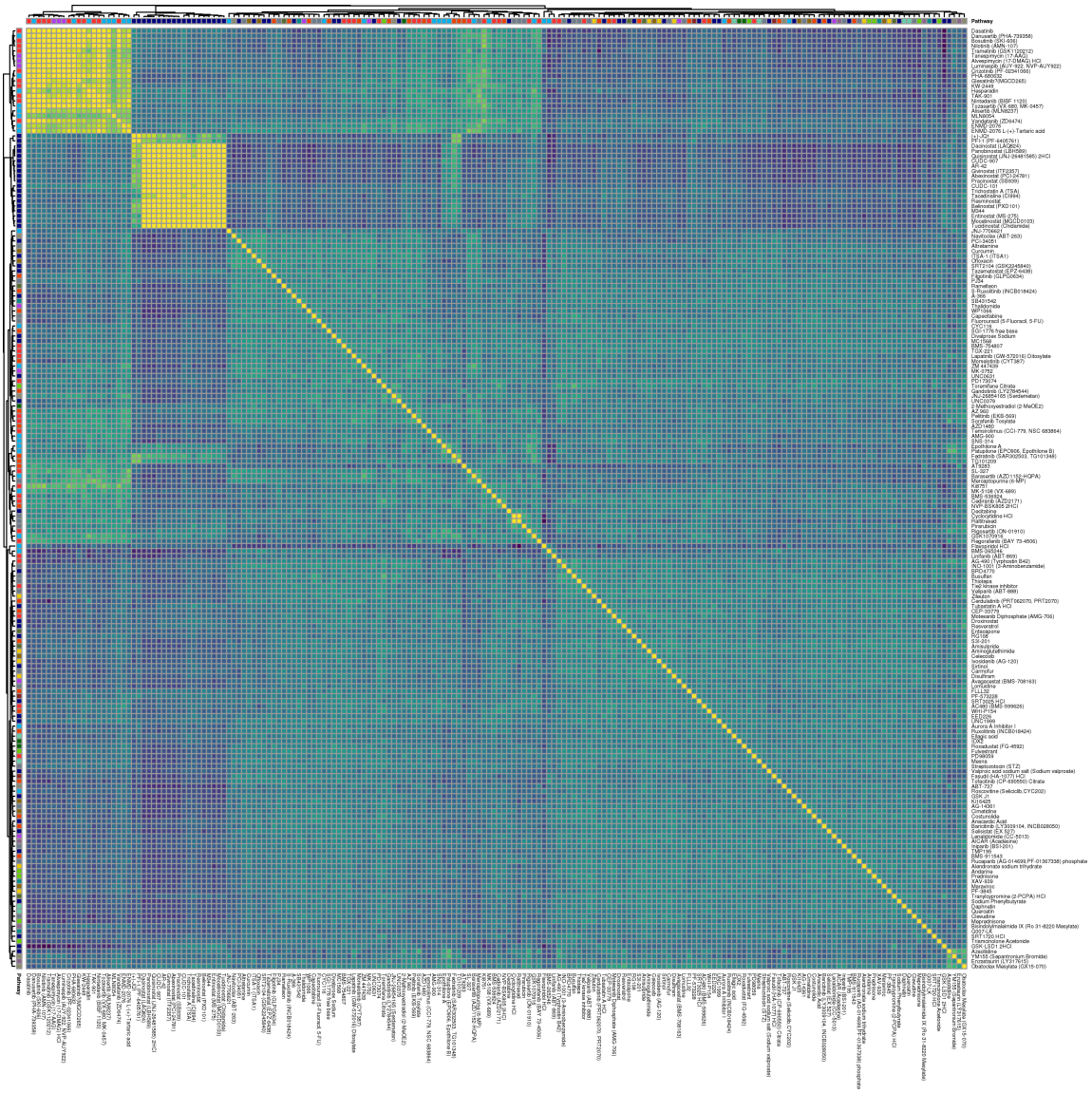

Figure 2.22: **Correlation of compound-driven molecular signatures for K562 cells identified in sci-Plex screen.** Heatmap depicts the Pearson correlation of beta coefficients across dose-dependent differentially expressed genes for every pairwise combination of compounds screened. To aid in visualization Pearson correlations were capped at 0.6.

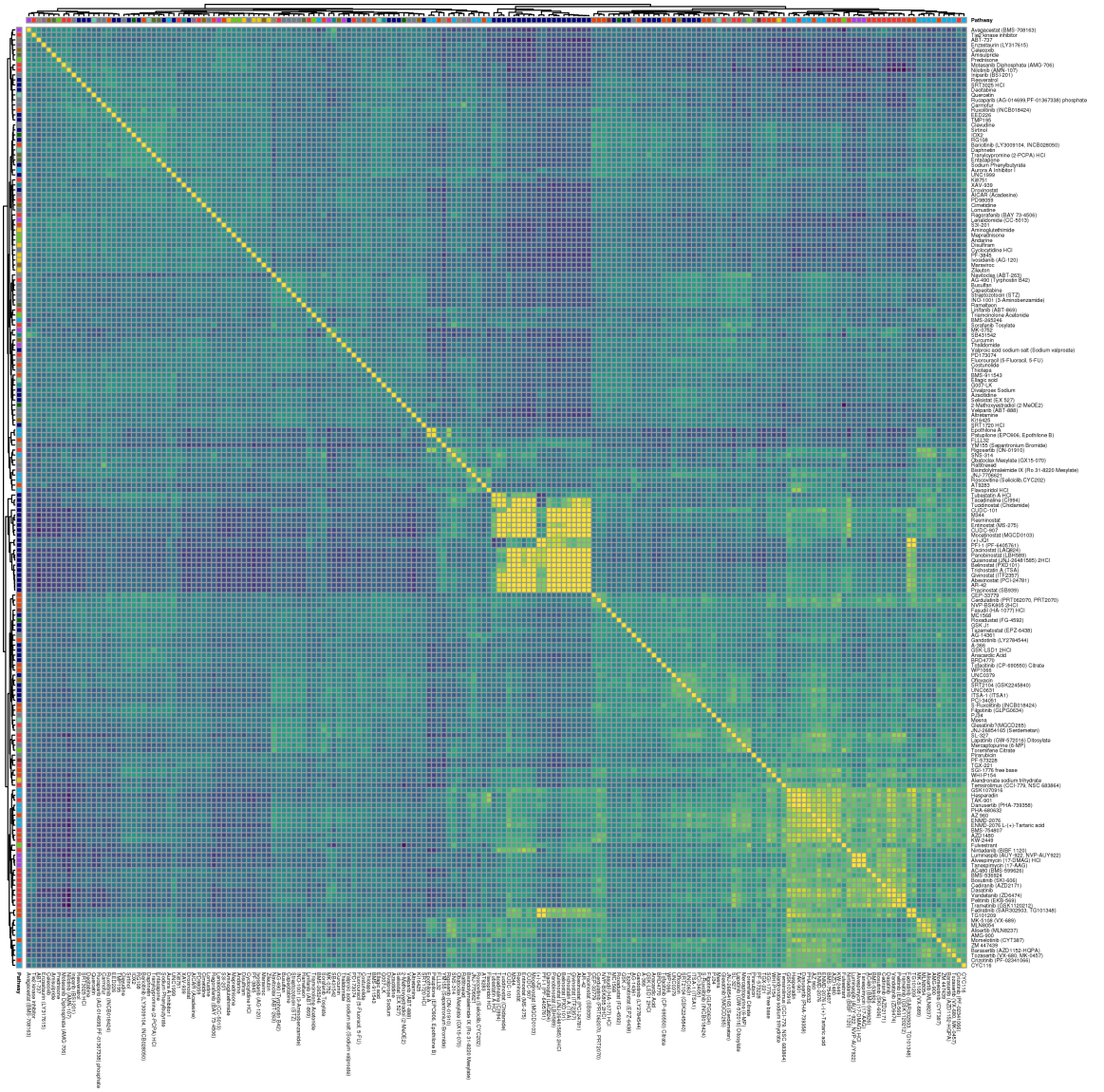

Figure 2.23: **Correlation of compound-driven molecular signatures for MCF7 cells identified in sci-Plex screen.** Heatmap depicts the Pearson correlation of beta coefficients across dose-dependent differentially expressed genes for every pairwise combination of compounds screened. To aid in visualization Pearson correlations were capped at 0.6.

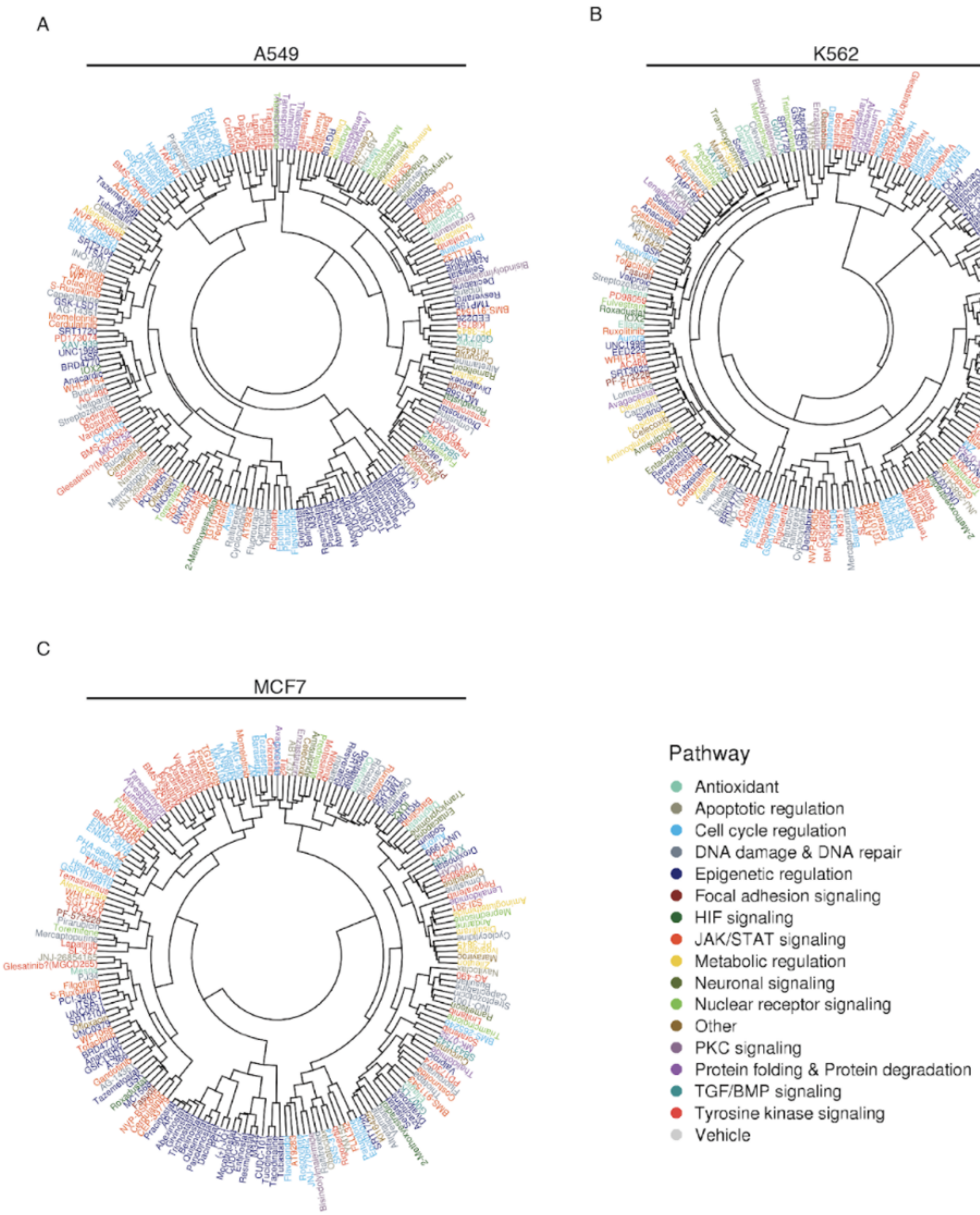

Figure 2.24: **Clustergrams of the correlation of compound-driven molecular signatures.** Clustergrams depicting the Pearson correlation of beta-coefficients across dose-dependent differentially expressed genes for every pairwise combination of compounds screened for A549 (A), K562 (B) and MCF7 (C) cells. Compounds names are colored by the pathway targeted.

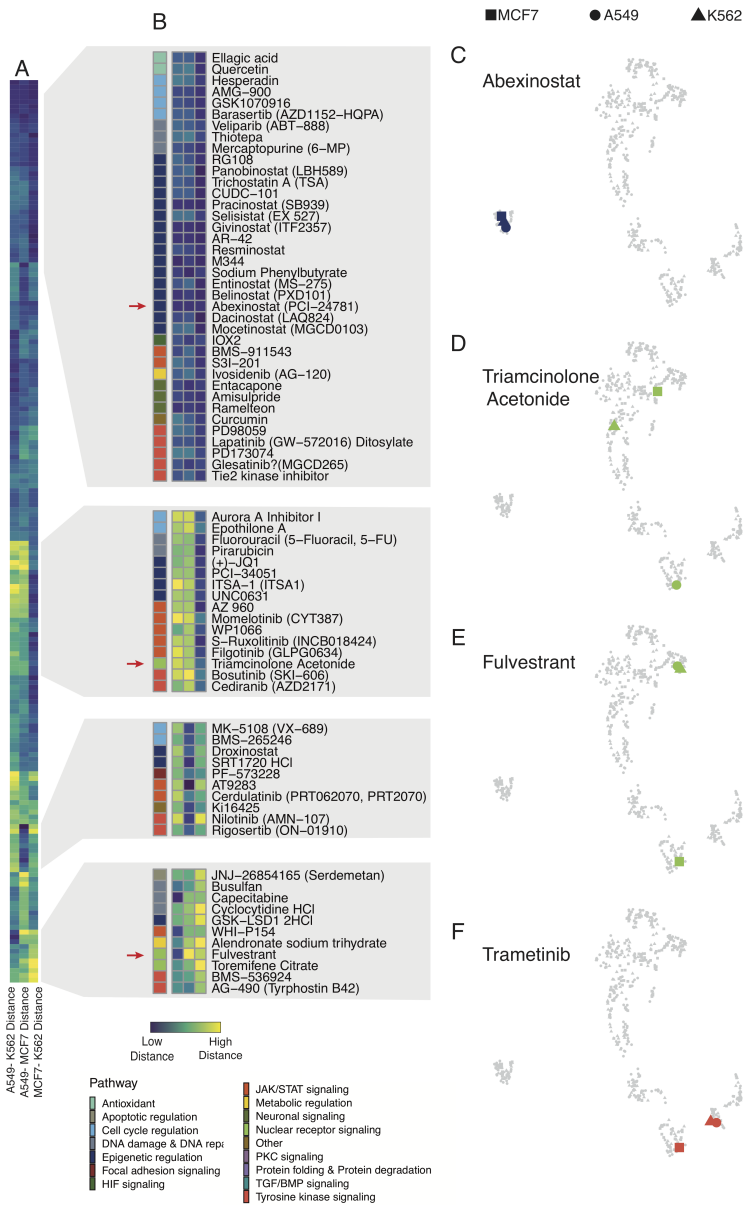

Figure 2.25: **Pairwise distances between PCA embeddings of drugs based on their dose-dependent effects.** A) Heatmap of pairwise distances between two cell types (columns) for a given drug (rows) in PCA reduced dimensional space. Hierarchically clustered to visualize cell type-specific responses to each drug. B) Insets of highlighted portions of the heatmap with pathway annotation shown to the left. Specific compounds highlighted with a red arrow are shown to the right (C-E) as UMAP embeddings. F) Trametinib treated cell lines are highlighted to illustrate colocalization of A549 and K562. Colored points correspond to labeled compound and all other drugs are shown in gray. Shape encodes the cell line from which each effect profile was captured (squares: MCF7; triangles: K562; circle: A549).

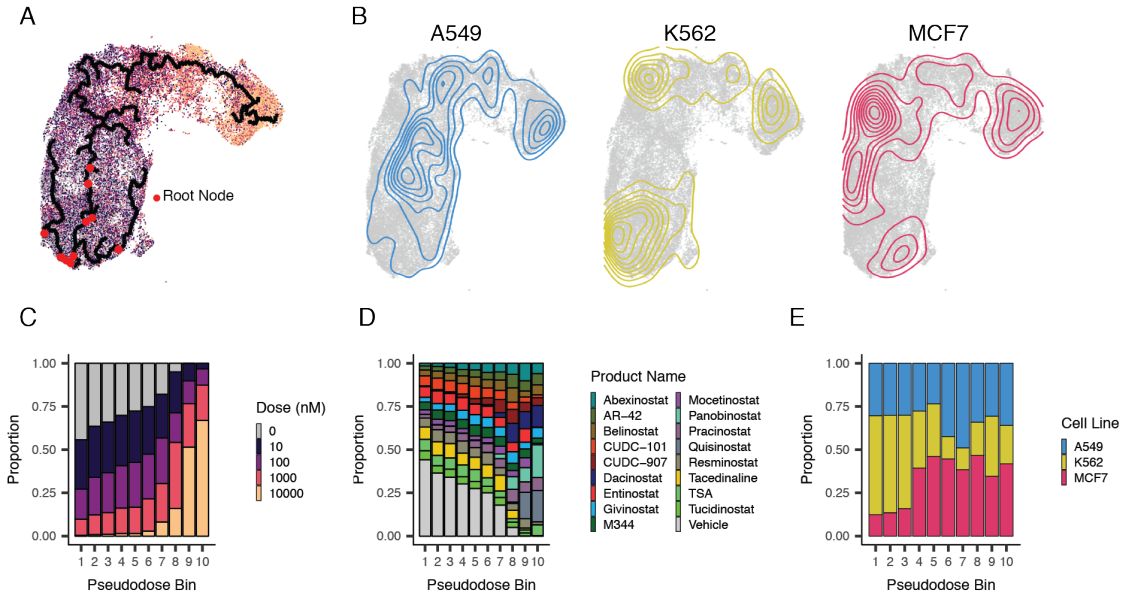

Figure 2.26: **HDAC inhibitor-treated cell types align and enable joint pseudodose trajectory reconstruction.**A) UMAP embedding highlighting the reconstructed pseudodose trajectory over the mutual nearest neighbor-aligned HDAC inhibitor and vehicle treated cells. Root nodes (red points) were chosen as nodes in the principal graph that had over 50% of their nearest neighbors annotated as vehicle treated cells. B) Distribution of each cell line within the embedding. C) Proportion within each pseudodose bin corresponding to each cell line.

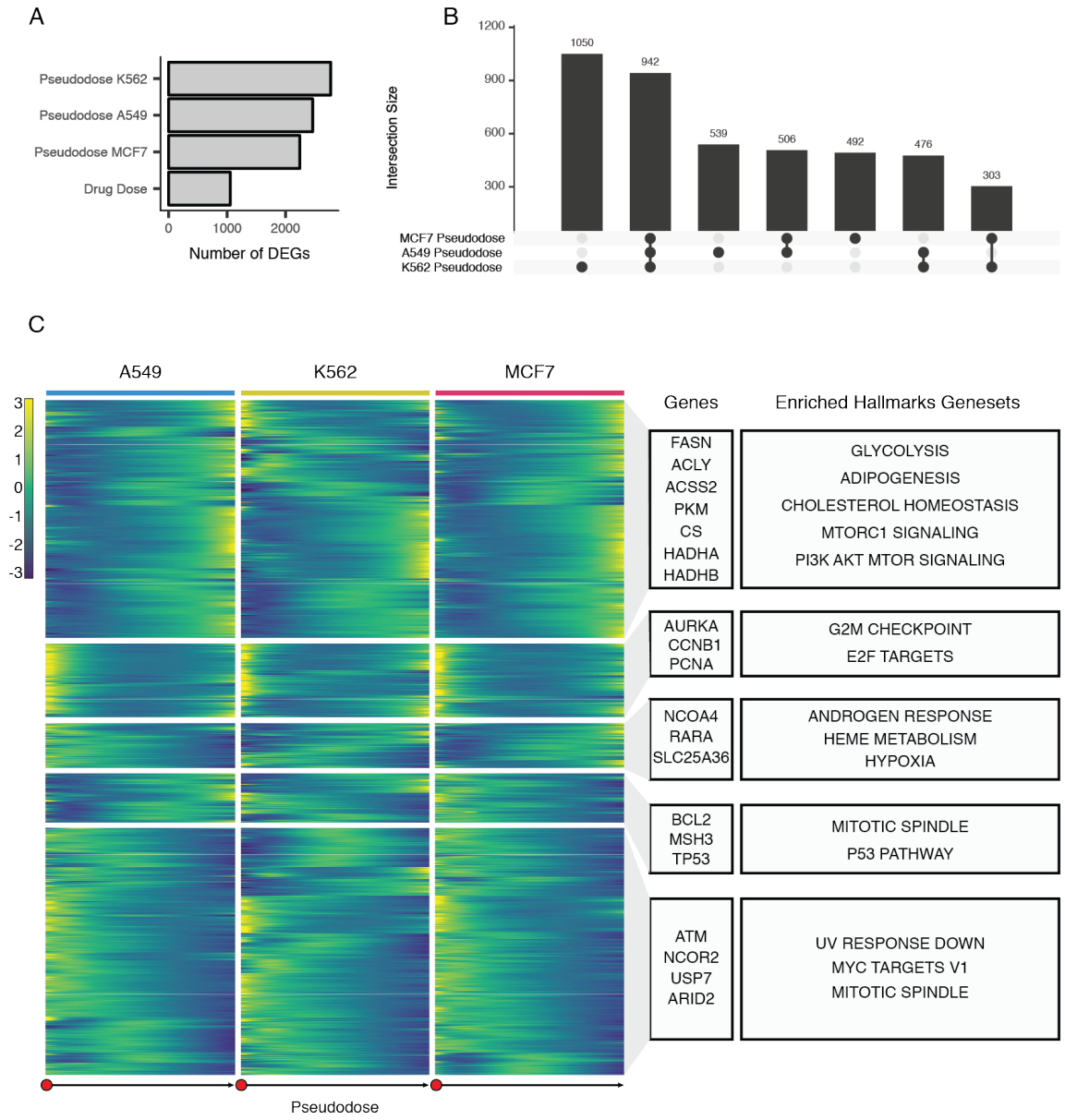

Figure 2.27: **Linear models identify pseudodose-dependent modules of proliferation and metabolism.** A) Barplot of the total number of significant dose-dependent and pseudodose-dependent DEGs (FDR $<$ 0.05). B) Upset plot displaying the intersections of significant pseudodose-dependent DEGs between the three cell types. C) Pseudodose heatmap depicting 4,308 genes that varied significantly as a function of pseudodose. Each row corresponds to the expected expression for a gene in the three cell lines as fit by the model described in the 'Differential expression analysis' section of the Methods. Genes (rows) were scaled and standardized within each cell line before joining the three matrices and performing hierarchical clustering. Clusters from hierarchical clustering were then used as an input into GSAhyper using the Hallmarks geneset. Select genes and genesets characterizing each cluster are shown (right).

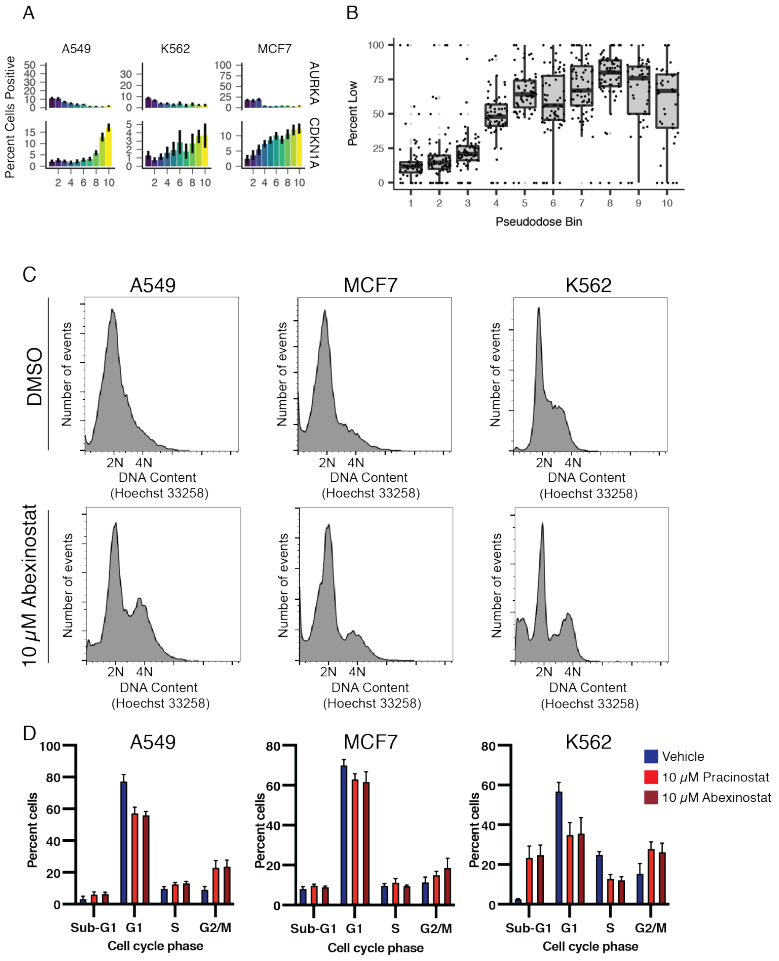

Figure 2.28: **HDAC inhibitor treatment induces cell cycle arrest in all three cell lines.** A) Percentage of cells expressing RNA for AURKA and CDKN1A across pseudodose bins. Black bars denote the bootstrapped 95% confidence interval. B) Boxplots depicting the percentage of cells in the low proliferation fraction in at a given drug dose across pseudodose bins. C) DNA content analysis of the three cell lines upon treatment with DMSO (top) or 10μM Abexinostat (bottom). D) Quantification of flow cytometry data depicting the number of cells in each DNA content category.

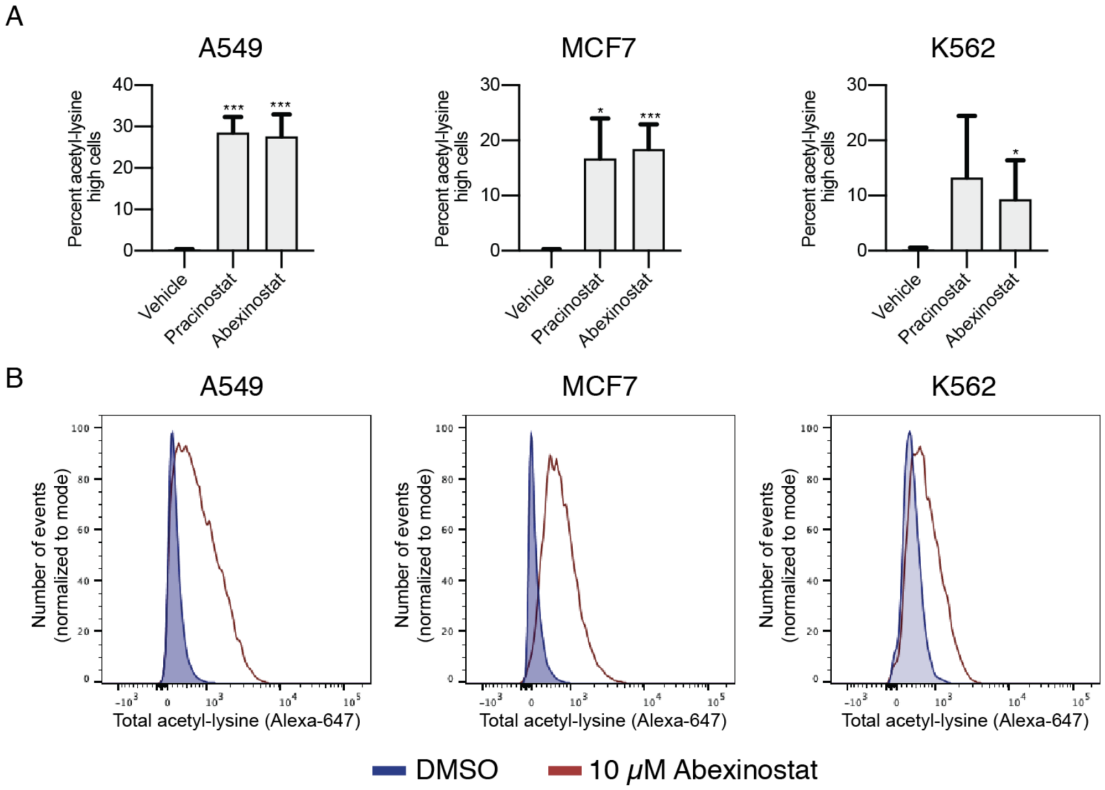

Figure 2.29: **HDAC inhibitor exposure leads to sequestration of acetate in the form of acetylated lysines.** A) Quantification of flow cytometry measurements of total cellular acetylated lysines in A549 (left panel), MCF7 (middle panel) and K562 (right panel) cells exposed to 10 μM pracinostat, 10 μM p abexinostat or vehicle control. Error bars denote standard deviation of the mean (Wilcoxon rank sum test, n = 3 culture replicates, * $p < 0.05$, *** $p < 0.005$). B) Representative flow cytometry histograms for the experiment quantified in panel A. Blue shaded regions and red lines correspond to DMSO vehicle control and 10 μM abexinostat, respectively.

Figure 2.30: **Pseudodose ridge plots.** Ridge plots display the distribution of cells along pseudodose for each HDAC inhibitor and dose combination for compounds that localized to the HDAC trajectory.

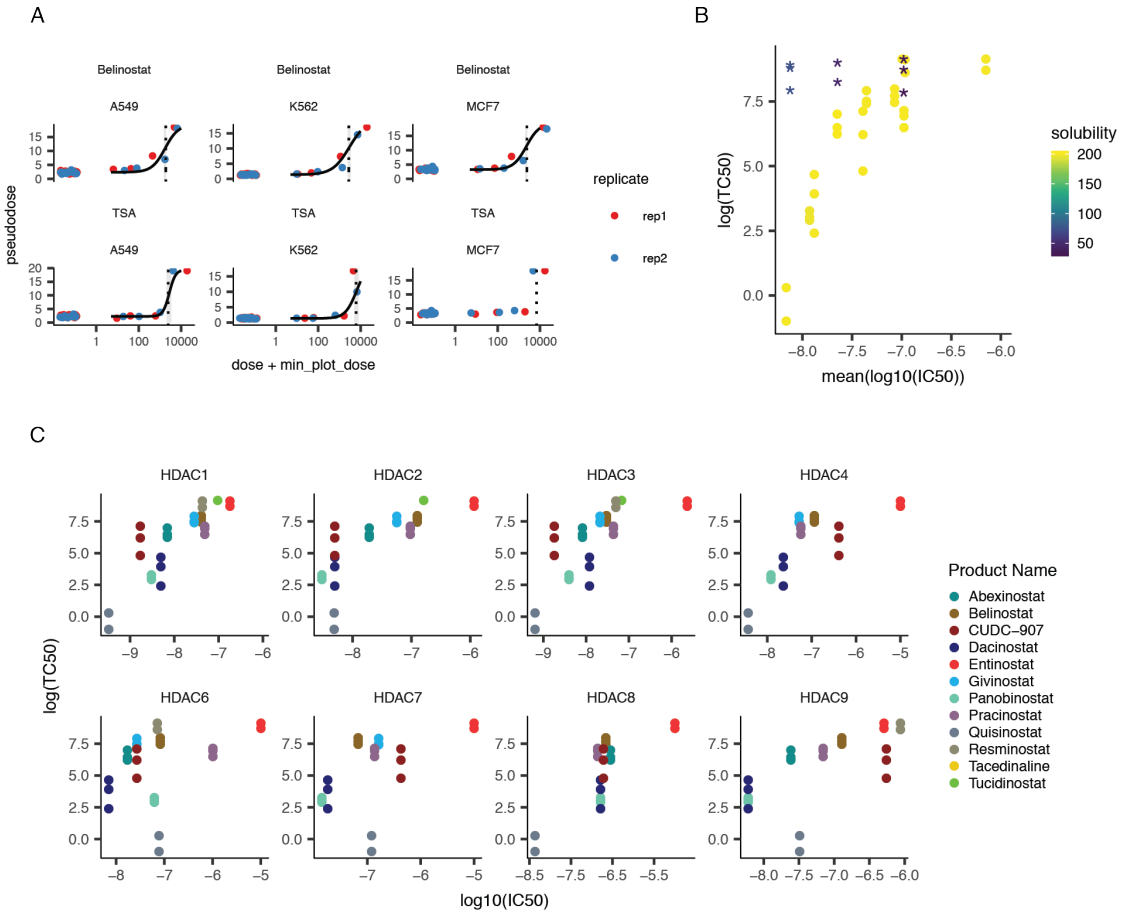

Figure 2.31: **Transcriptional trajectory of HDAC inhibitor-treated cells corresponds to in vitro IC50 measurements.**A) Pseudodose response curves were fit for each compound and each cell line using the drc R package. The mean position of each dose along the pseudodose trajectory was used as the response. Two illustrative examples for Belinostat (top) and Trichostatin A (TSA) (below) are shown. Dotted vertical lines illustrate the transcriptional EC50 (TC50) for each compound in each cell line. Shaded gray area denotes the 95% confidence intervals for each TC50 estimate. B) Plot displaying aggregate in vitro measured mean of log10(IC50 [M]) versus log(TC50) colored by solubility supplied by Selleckchem Chemicals. Points displayed as (*) were not used for fits. C) log10(IC50 [M]) versus log(TC50) for each HDAC isoform. Each point is colored by the HDAC inhibitor used.

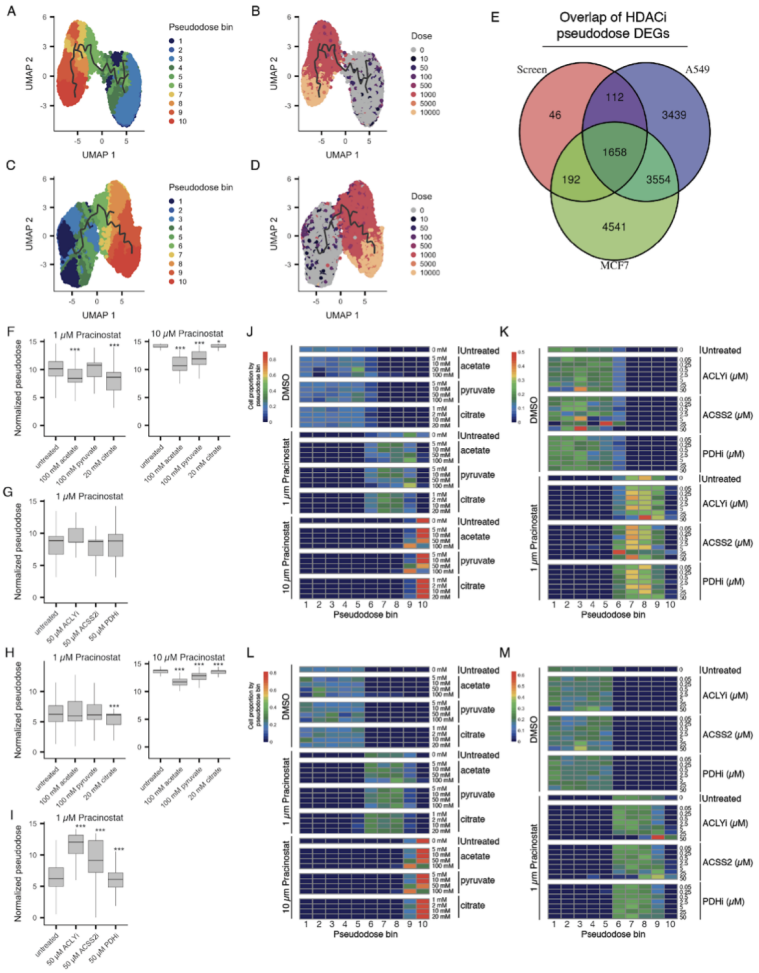

Figure 2.32: **Supplementation with acetyl-CoA precursors decrease, while inhibition of enzymes that replenish acetyl-coA pools exacerbate, progression along the HDAC inhibitor pseudodose trajectory.**A-D) UMAP embeddings of A549 (A,B) and MCF7 (C, D) single cell transcriptomes after exposure to the HDAC inhibitors pracinostat or abexinostat, in the presence or absence of acetyl-CoA precursors or inhibitors to enzymes that replenish acetyl-CoA pools. UMAP were constructed from cells from all conditions in the experiment. Cells are colored by pseudodose bin (A,C) or dose (B,D). E) Venn diagram of the overlap of differentially expressed genes across trajectories between or original HDACi trajectory vs. A549 or MCF7 HDACi trajectories from this new experiment. F,H) Boxplots of pseudodose estimates for select conditions of cells exposed to 1 or 10 μM pracinostat with or without co-treatment with acetyl-coA precursors for A549 (H) or MCF7 (L) cells. Values are normalized to vehicle treated cells. Wilcoxon rank sum test. G,I) Boxplots of pseudodose estimates for select conditions of cells exposed to vehicle and pracinostat with or without co-treatment with acetyl-coA precursors for A549 (I) or MCF7 (M) cells. Values were normalized to vehicle treated cells. Wilcoxon rank sum test. J,L) Heatmaps depicting the fraction of cells per pseudodose bin for cells exposed to various acetyl-coA precursors in pracinostat-exposed A549 (F) or MCF7 (J) cell. K,M) Heatmaps depicting the fraction of cells per pseudodose bin for cells exposed to various inhibitors targeting enzymes that replenish acetyl-coA pools in pracinostat-exposed A549 (G) and MCF7 (K) cells.

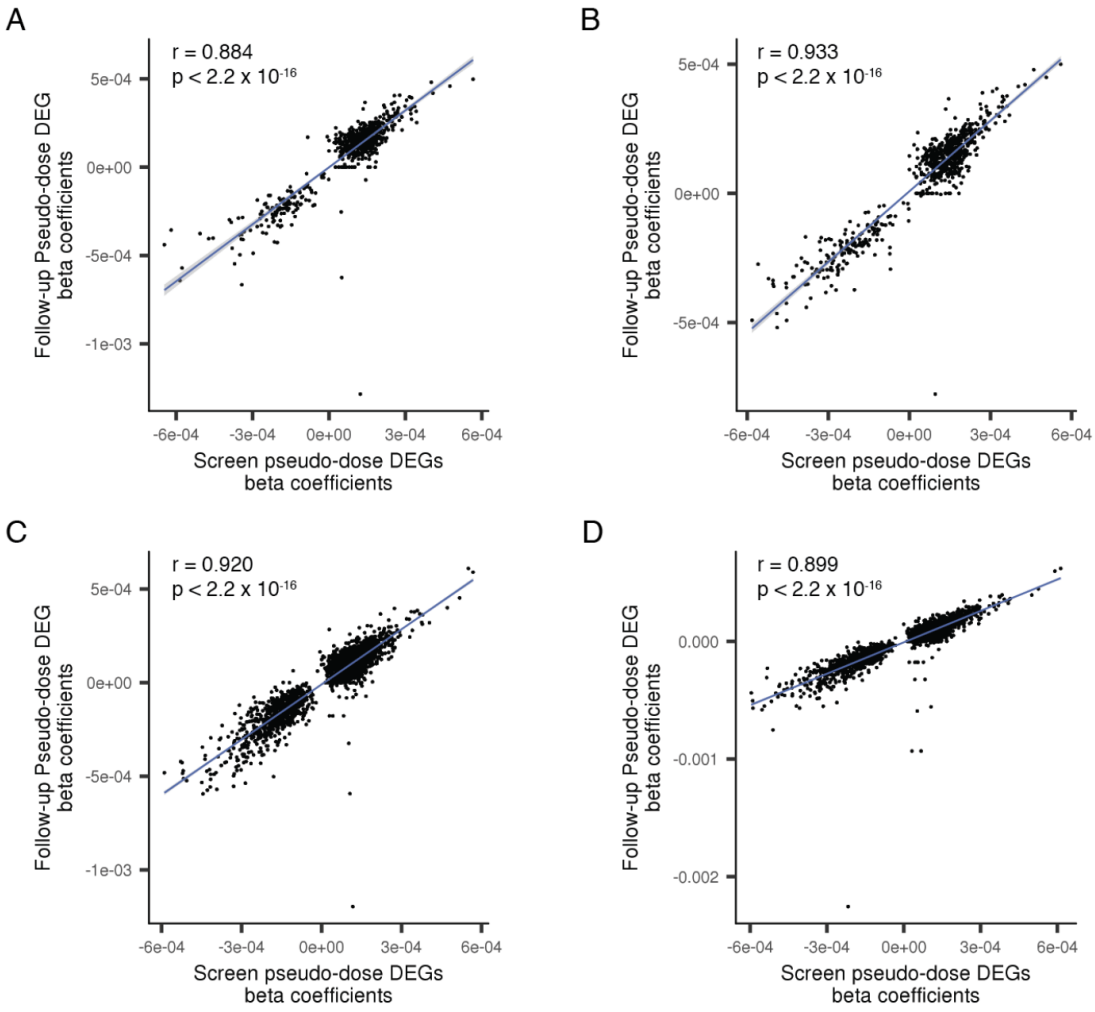

Figure 2.33: **Correlation of effect sizes between differentially expressed genes post-HDAC inhibition from original screen vs. new experiment.** A-B) Correlation of effect size estimates (beta coefficients) for differentially expressed genes between vehicle control and 10 μM abexinostat (A) or 10 μM pracinostat (B) for A549 cells. C-D) Correlation of effect size estimates (beta coefficients) for differentially expressed genes between vehicle control and 10 μM abexinostat (C) or 10 μM pracinostat (D) for MCF7 cells. X-axes correspond to large-scale sci-Plex experiment. Y-axes correspond to targeted followup sci-Plex experiment.

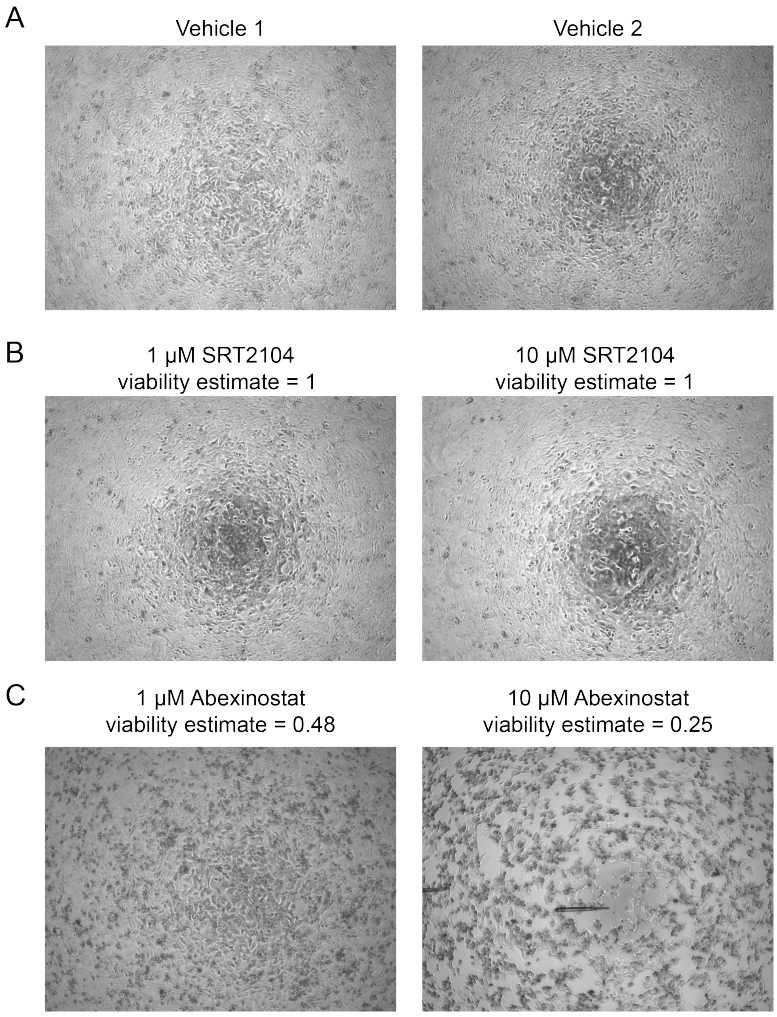

Figure 2.34: **Contact inhibition of cell proliferation 72 hours post drug exposure.** Representative bright-field images of A549 cells exposed to vehicle (A) or the specified dose of the SIRT1 activator SRT2104 (B) or the HDAC inhibitor Abexinostat (C). Viability estimates as determined by recovered cell counts for each drug/dose combination normalized to cell counts of vehicle control wells.

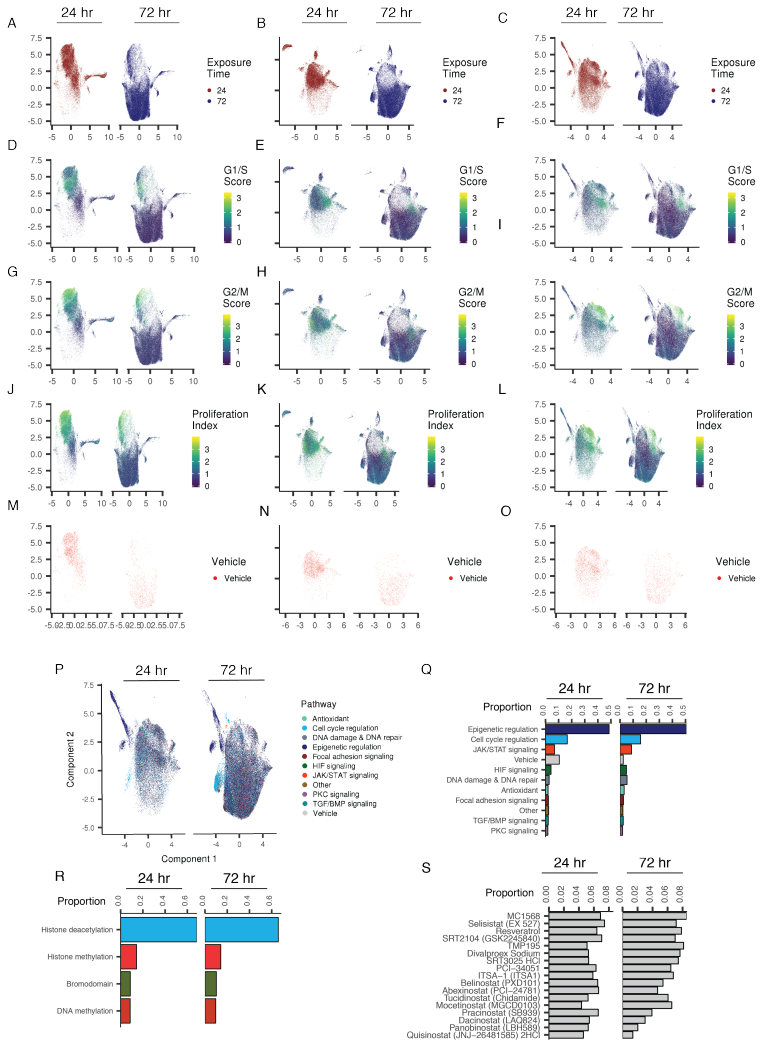

Figure 2.35: **Aligning A549 cells at 24 and 72 hours after treatment reveals time-dependent responses to diverse small molecules.**A-C) UMAP embedding of A549 cells at 24 and 72 hours post treatment in the absence of a correction for differences in viability and proliferation (A), after linear transformation of the data to account for changes in proliferation index and viability (B) and after mutual nearest neighbor based alignment of data after linear transformation (C). Cells are colored by the time point at which they were collected. (D-F) UMAP embeddings as in panels A-C with cells colored by the aggregated normalized expression score of G1/S marker genes. (G-I) UMAP embeddings as in panels A-C with cells colored by the aggregated normalized expression score of G2/M marker genes. (J-L) UMAP embeddings as in panels A-C with cells colored by proliferation index. (M-O) UMAP embeddings as in panels A-C only visualizing cells treated with vehicle control. (P) UMAP embeddings from panel C with cells colored as to the pathway targeted by the treatment to which they were exposed. (Q) Proportion of cells broken up by pathway targeted. Note that only a subset of our 188 compounds across a limited number of pathways were tested at 72 hours. (R) Proportion of cells broken up by the activity targeted by treatment with epigenetic regulation compounds. (S) Proportion of cells broken up by HDAC compound.

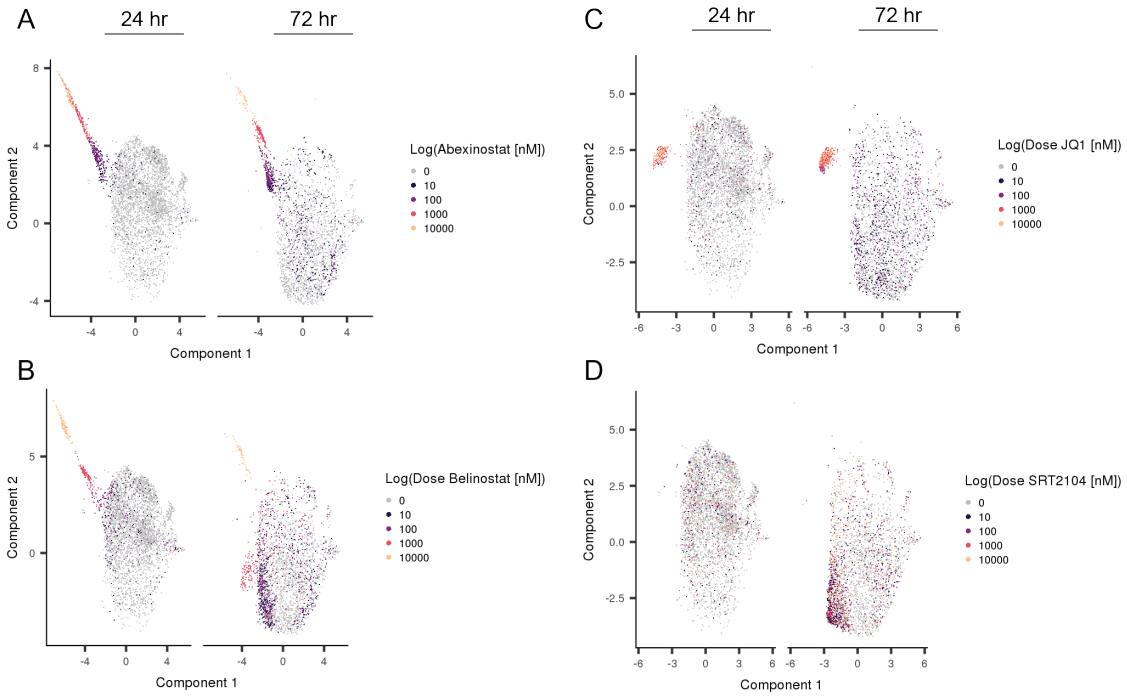

Figure 2.36: **Bromodomain inhibition, sirtuin activation, and histone deacetylase inhibition induce characteristic transcriptomic responses.**A-D) UMAP embedding of MNN aligned A549 cells 24 and 72 hours after treatment with the pan-HDAC inhibitors abexinostat (A) or belinostat (B), the bromodomain inhibitor JQ1 (C), and the SIRT1 activator SRT2104 (D). Cells are colored by the dose to which each cell was exposed.

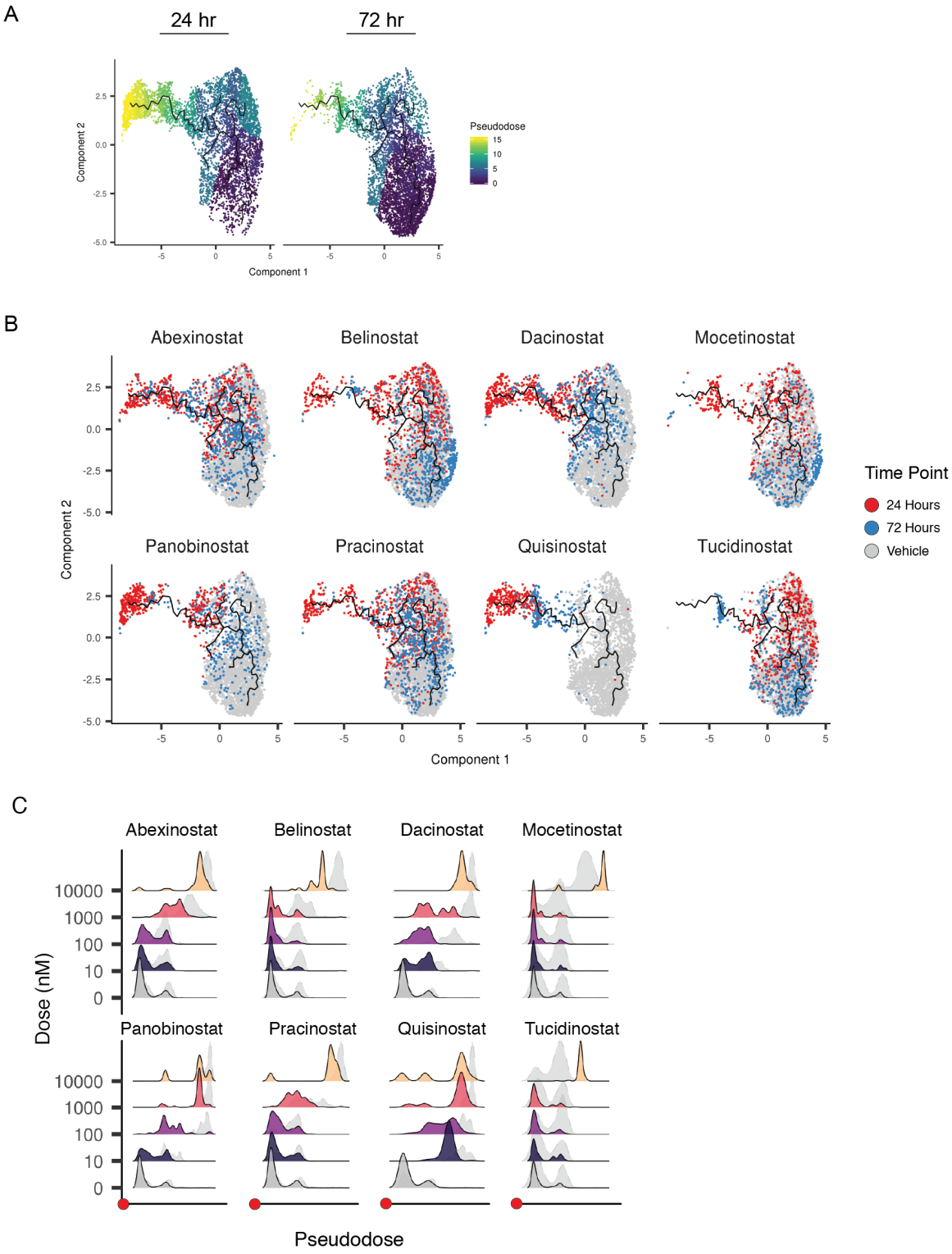

Figure 2.37: **The heterogeneous response to the majority of HDAC inhibitors does not appear to be driven by cellular asynchrony.** A) Aligned UMAP embeddings of cells exposed to vehicle HDAC inhibitors for 24 or 72 hours. Cells are colored by their progression along pseudodose. B) Aligned UMAP embeddings of cells exposed to vehicle (grey cells) or the labeled HDAC inhibitor for 24 (red cells) or 72 (blue cells) hours. C) Pseudodose position of cells upon treatment with the indicated HDAC inhibitor. 72 hour time point is shown in color, while 24 hours is shown in transparent gray.

Chapter 3: EMBRYO-SCALE, SINGLE CELL SPATIAL TRANSCRIPTOMICS

Chapter 3 is adapted with minimal modification from:

Sanjay R. Srivatsan*, Mary C. Regier*, Eliza Barkan, Jennifer M. Franks, Jonathan S. Packer, Parker Grosjean, Madeleine Duran, Sarah Saxton, Jon J Ladd, Malte Spielmann, Carlos Lois, Paul D. Lampe, Jay Shendure, Kelly R. Stevens, Cole Trapnell *Science* 2021(in press).

I contributed to all aspects of the paper, including the conception of the project, performance of the experiments, analysis of the data, and writing the initial manuscript draft.

3.1 ABSTRACT

Spatial patterns of gene expression span many scales, and are shaped by both local (e.g. cell-cell interactions) and global (e.g. tissue, organ) context. However, most in situ methods for profiling gene expression either average local contexts or are restricted to limited fields of view. Here we introduce sci-Space, a scale-flexible method that retains single cell resolution while resolving spatial heterogeneity in gene expression at larger scales. As a proof-of-concept, we apply sci-Space to whole mount sections of developing mouse embryos (E14.0), capturing the approximate spatial coordinates and whole transcriptomes of 120,000 profiled nuclei. We identify thousands of genes that are expressed in an anatomically patterned manner, including both known (e.g. Hox genes) and novel factors, and find that individual cell types vary substantially in the extent to which they exhibit spatial patterning of gene expression at this scale. In both the nervous system and mesenchyme, spatial information facilitates the annotation of cellular subtypes. Finally, in the developing mouse brain, we identify strong correlations between pseudotime and three spatially distinct migratory patterns of differentiating neurons. Looking forward, sci-Space presents a path for the construction of spatially-resolved single cell atlases of mammalian development.

3.2 INTRODUCTION

Methods for molecular profiling at single cell resolution have the potential to transform our understanding of developmental biology. For example, in diverse models of embryogenesis, we and others have performed "whole organism" profiling of transcription or chromatin accessibility at single cell resolution (*84–90*). Although the clear majority of major cell types have previously been described, such studies yield a much richer view of cell states during development than was previously available. For example, applying machine learning techniques to organize cells in terms of their progression through development has revealed complex branching trajectories similar to, but topologically distinct from, cell lineage histories (*91*, *92*).

The spatial organization of cells plays a central role in normal development and homeostasis, as well as in the pathophysiology of many diseases. However, a key limitation of widely used methods for single cell molecular profiling is that they operate on disaggregated cells or nuclei. Although several groups have developed powerful in situ methods for measuring the expression of many or all genes while retaining spatial information, these methods are not without limitations **(Figure 3.6)**. A first class of methods, including the original "spatial transcriptomics" (*93*) and Slide-seq (*94*) methods, barcode and then count mRNAs derived from each spot across patterned arrays. However, although such methods can potentially be implemented at a range of spatial scales, a key limitation is that the boundaries of spots have no natural correspondence to the boundaries of cells. As such, they yield aggregate profiles of small regions encompassing multiple cells and/or portions of cells, rather than resolving individual cells. A second class of methods, including MERFISH (*95*), seqFISH (*96*), and FISSEQ (*97*), rely on in situ hybridization or sequencing to measure the expression of many genes while retaining single cell (or even subcellular) resolution within each field of view. However, such multiplex in situ methods are typically limited by long image acquisition times and complex instrumentation requirements. In sum, the tradeoffs between these two classes of techniques is such that assaying the whole transcriptomes of individual cells over large regions remains impractical.

We sought to develop a spatial transcriptomics method that retained both single cell resolution

(that is, the ability to unambiguously ascertain sets of transcripts expressed in the same cell) and the flexibility to capture positional information at broader scales (e.g. to detect spatial patterns of cell type-specific gene expression that would be found at the level of a whole organ or embryo, but missed by high-resolution analysis of a small region). To this end, we hypothesized that we could label cells with molecular tags that encoded their approximate spatial coordinates within entire tissue sections, and subsequently recover this information without sacrificing single cell resolution. Furthermore, sampling single cells from across a tissue could reveal spatial patterns of gene expression, including those that are cell type-specific, without comprehensively (and expensively) sequencing the whole transcriptome of every cell.

3.3 RESULTS

3.3.1 Spatial labeling of nuclei with oligonucleotide hashes

We recently developed sci-Plex, a method for labeling or "hashing" nuclei using unmodified DNA oligos during single cell RNA-seq with combinatorial indexing (sci-RNA-seq) (*98*). sci-Plex enables the pooling of nuclei from many different specimens or samples into one sci-RNA-seq experiment at minimal marginal cost. To leverage parts of this workflow to capture spatial information, we reasoned that we could first print unique combinations of hashing oligos to a spatially patterned array, and then transfer those oligos to the nuclei of a tissue slice by diffusion (*99*). Provided that these hashing oligos were recovered in association with single cell RNA-seq profiles, they could be used to reconstruct each cell's approximate tissue coordinates upon sequencing.

As a proof-of-concept of this method, which we call "sci-Space", hashing oligos were spotted onto glass slides coated with a thin layer of dried agarose in a gridded format. These spatial grids contained 7,056 uniquely barcoded spots in a 18mm by 18mm grid with a mean radius of 73.2 $\pm$ 14.1μm and a mean spot-to-spot center distance of 222 $\pm$ 7.5μm **(Figure 3.7)**. About 5% of spots, constituting an identifiable pattern, were also loaded with SYBR green fluorescent dye. After transferring the oligos from the slide to the tissue, the grid could be registered with an image of the tissue using these concurrently imaged fluorescent fiduciary points **(Figure 3.7, Figure 3.8)**.

Next, we optimized hash oligo concentrations to robustly label nuclei from sectioned tissue, developed a protocol for blotting the hash-oligos onto the tissue, and verified that we could dissociate, recover, and sequence single cell transcriptomes from labeled nuclei from cultured cell lines or sectioned tissue affixed to a glass slide **(Figure 3.9, Figure 3.10)**. To maximize the recovery of cells sequenced from each slide, we optimized our protocol to thoroughly dissociate cells, and altered the procedure for seeding sci-RNA-seq wells (FACS $\rightarrow$ dilution). After these optimizations, we arrived at the sci-Space protocol, which consists of four steps: 1) fresh-frozen tissue is sectioned; 2) sectioned tissue is permeabilized and physically juxtaposed to a glass slide bearing the spatially gridded hashing oligos; 3) during oligo transfer, the spatial grid is imaged; and 4) nuclei from the tissue on the slide are extracted, further hashed with a slide-specific oligo, chemically fixed and subjected to sci-RNA-seq **(Figure 3.1A**; **Figure 3.11, Figure 3.12)**.

3.3.2 Spatially-resolved single cell sequencing of the mouse embryo

To apply sci-Space, we profiled fourteen sagittal sections derived from two E14.0 mouse embryos (C57BL/6N). After sequencing, conservative doublet removal (*100*), quality filtering by RNA UMIs, and assignment of each cell to a slide (based on its slide-specific hash oligo), our dataset comprised 121,909 spatially resolved single cell transcriptomes, with an average of 2,514 unique molecular identifiers (UMIs) per cell, 1,231 genes detected per cell, and without apparent batch effects between slides or embryos **(Figure 3.13, Figure 3.14)**. This corresponds to capture of 164 nuclei/mm2 of tissue on average, or sampling of 2.2% of the estimated nuclei present **(Figure 3.15)**. Rather than annotating cell types ab initio, we co-embedded (*101*, *102*) these data with published, non-spatial sci-RNA-seq "mouse organogenesis cell atlas" (MOCA) data spanning E9.5 to E13.5 (*84*). Reassuringly, these E14.0 data distributed to the distal tips of various E9.5 to E13.5 trajectories (**Figure 3.1B**). An initial set of cell type annotations were inferred by nearest neighbor label transfer between the E13.5 time point of the MOCA dataset and the sci-Space dataset. We also transferred annotations from a developing mouse brain atlas (DMBA) spanning E13.5 to E14.5 (*103*). The cell type annotations inferred by nearest neighbor label transfer were highly concordant

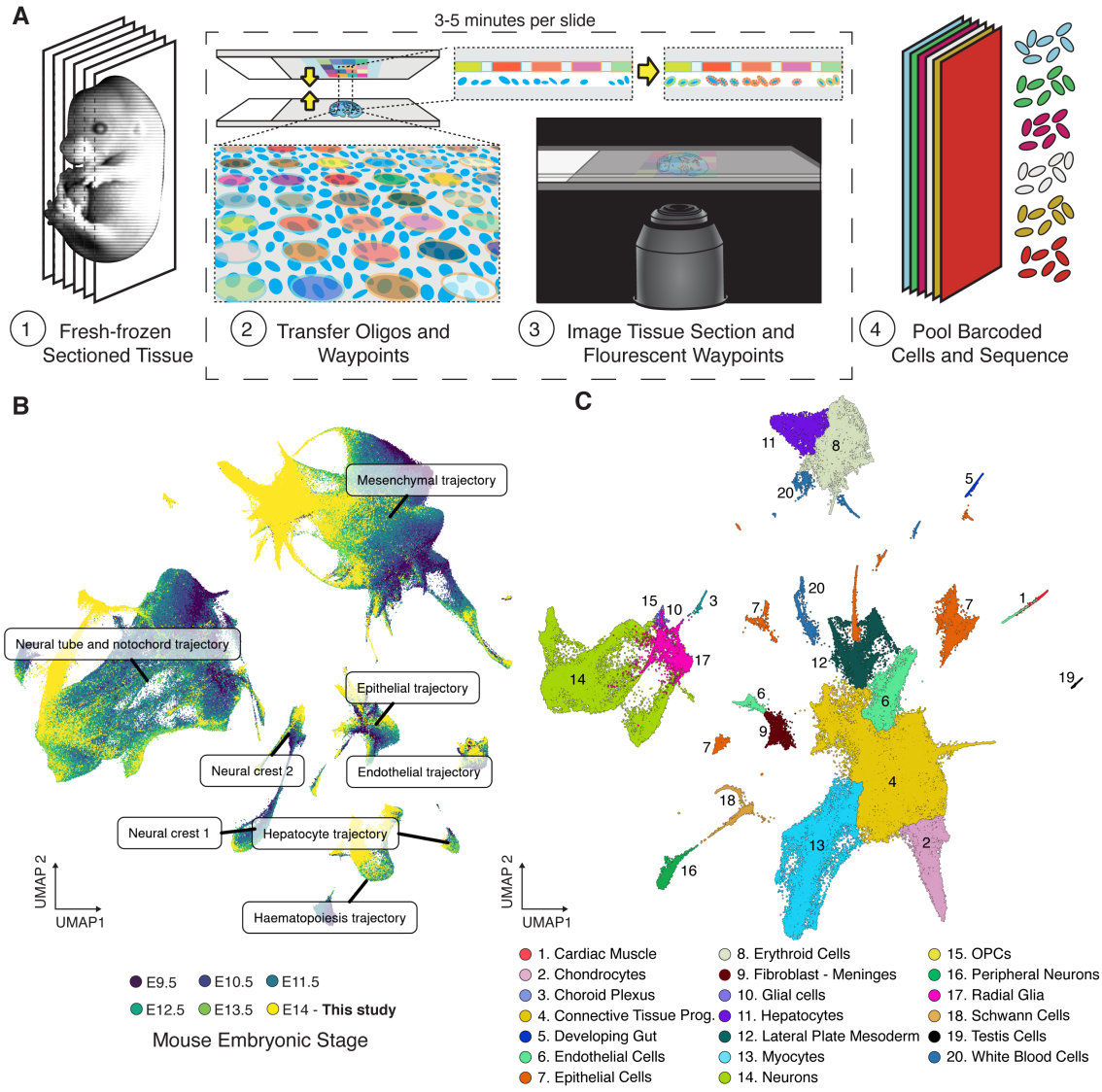

Figure 3.1: **sci-Space recovers single cell transcriptomes while recording spatial coordinates.** (**A**)Fresh-frozen, sectioned tissues are exposed to a grid of arrayed single-stranded oligos, a subset of which are fluorescent waypoints, and then imaged. The oligo transfer and imaging procedure takes 3 to 5 minutes per slide. The cells from each slide are labeled with an additional, section-identifying barcode prior to pooling and sci-RNA-seq. (**B**) Joint embedding of E14.0 single cell transcriptomes from this study (yellow) and published data from single cell transcriptomes spanning E9.5 to E13.5 (*84*). Major trajectories are labeled. (**C**) UMAP embedding of 121,909 cells from sectioned E14.0 mouse embryos. Cell types are denoted by color and number.

with those derived from an orthogonal approach using Garnett (*104*), a semi-supervised method for annotating single cell data **(Figure 3.16)**. Finally, these annotations were refined by manual inspection of differentially expressed genes (**Figure 3.1C**).

Images of sectioned embryos and sequencing data were co-registered by matching the SYBR waypoints in the oligo grid to identifiable spots of SYBR staining in the image (*105*, *106*) **(Figure 3.2A**; **Figure 3.17, Figure 3.18)**. This alignment yielded an affine transformation from grid positions to the image, from which cell coordinates were then calculated within the sectioned tissue **(Figure 3.18)**. Nuclei were mapped to the position with the highest combination of spot and sector oligos within the imaged section. For about 9% of nuclei, the top assignment was not located near any other nuclei of the same cell type; for such "outliers", alternative mappings were considered (Methods). Altogether, 121,909 nuclei were well-localized **(Figure 3.13C-F)** to one of 15,102 spatial positions across 14 sections; on average, each spatial position was assigned of 8.1 nuclei (10.5 s.d.) that when summed, included 20,296 UMIs (29,465 s.d) and 4,428 genes (3,251 s.d) **(Figure 3.19)**.

We next quantified the distribution of cell type annotations across each section. First, each slide's image was segmented into readily discernible organs **(Figure 3.2B; Figure 3.20)**, a process aided by immunostaining and alignment of adjacent sections (*107*) **(Figure 3.21)**. As expected, neurons mapped largely within the spinal cord and brain outlined by cells of the developing meninges, cardiomyocytes within the heart, and white blood cells throughout the organism **(Figure 3.2C; Figure 3.22)**. Although several cell types mapped almost exclusively to one anatomic segment (e.g. hepatocytes to liver), others mapped more broadly (e.g. connective tissue progenitors) **(Figure 3.2D; Figure 3.22, Figure 3.35)**. Analysis with Giotto (*108*), an unsupervised tool for segmenting spatial transcriptomic images into tissue "domains" of similar cell type composition, revealed 22 domains shared across sci-Space slides. In addition to detecting boundaries between major organs, Giotto was able to automatically recognize more complex domains with distributions extending throughout the embryo (e.g. mesenchymal tissue and cartilage) **(Figure 3.36)**.

Finally, wherever cells are captured, sci-Space data enables the visualization of any gene in the transcriptome akin to an in situ hybridization, albeit at lower spatial resolution. For example, a

sci-Space "digital in situ" of the dopamine transporter *Slc6a3* highlights a cluster of dopaminergic neurons at the midbrain-hindbrain boundary, consistent with stage- and section- matched whole-mount in situs (**Figure 3.2E**); additional such examples are shown in **Figure 3.37**. However, unlike conventional in situ data, sci-Space data also resolves gene expression attributable to different cell types across the embryo. For example, in the heart, both cardiomyocytes and endothelial cells in the heart express the growth factor *Fgf1*, while only cardiomyocytes express the growth factor receptors *Fgfr1*, *Fgfr2* and *Fgfr3* (**Figure 3.2F**).

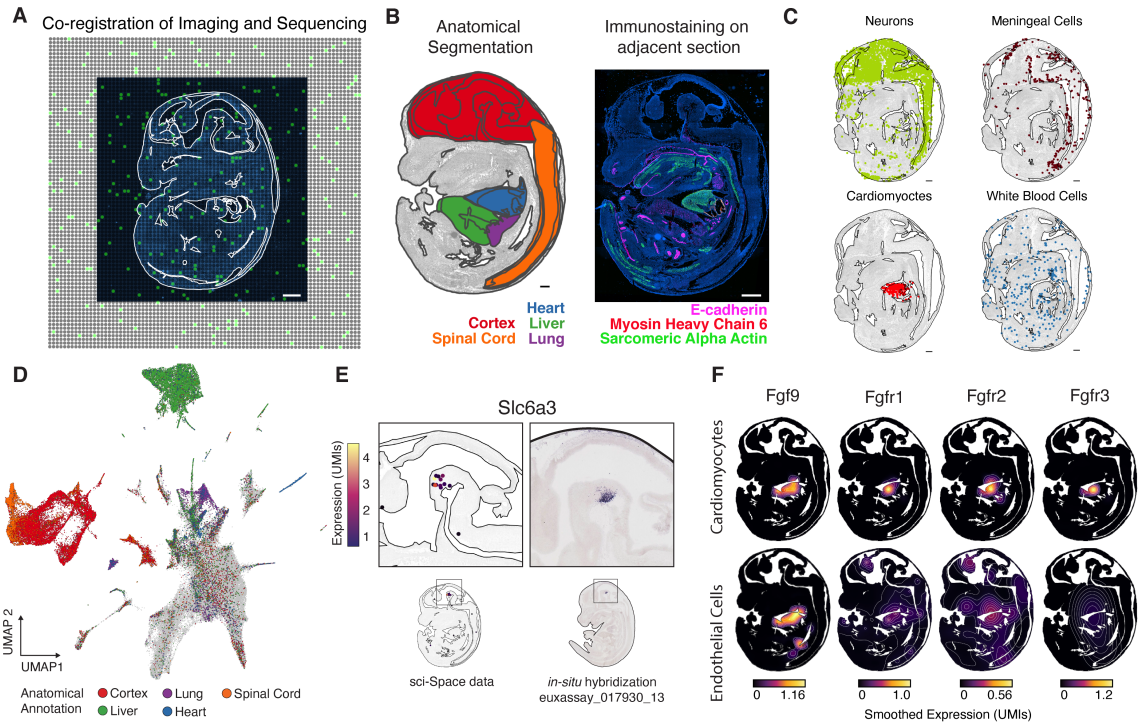

Figure 3.2: **sci-Space captures spatially and cell type resolved gene expression across the embryo.** (**A**) Co-registered DAPI stained section image and oligo array, superimposed. SYBR waypoints are highlighted as green spots. (**B**) Anatomical regions of Slide 1 (left) and an adjacent immunostained serial section aligned to Slide 1 (right). (**C**) Highlighted cell types mapping to a single slide. (**D**) UMAP embedding colored by anatomical regions. (**E**) Gene expression of dopamine transporter Slc6a3 from sci-Space data (left) and published (*109*) section/stage matched in situ (right). (**F**) Smoothed percentage of gene expression for Fgf1, Fgfr1, Fgfr2 and Fgfr3 in either cardiomyocytes (top) or endothelial cells (bottom). Color scaled to maximum percentage within each gene. Scale bars in panels (**A-C**) = 0.5 mm.

3.3.3 Contrasting sci-Space and spatial transcriptomics methods relying on patterned arrays.

A key distinction between sci-Space and STC (spatial transcript capture; **Figure 3.6, left**) such as Slide-seq, is that because STC methods capture transcripts from lysed tissue sections (*93*, *94*), each spot can include RNA from multiple cells and/or portions of cells. As such, STC methods are limited in their ability to resolve gene expression variation within individual cells or cell types. To quantify the consequences of this limitation, we aggregated transcriptomes from sci-Space cells derived from the same spatial positions, effectively creating a "mock-STC" dataset. Cell types and clusters readily discernible in the single cell data were hard to resolve or lost entirely in the mock-STC data **Figure 3.38A-C)**. One approach to ameliorate this limitation involves mapping cells profiled by non-spatial single-cell RNA-seq onto the spatial "scaffolds" afforded by STC, thereby imputing cellular locations within a tissue (*102*, *110*). To assess the viability of this approach, we performed such imputation on the mock-STC data, and compared it to the single nucleus data used to derive it **Figure 3.38D,E)** (*102*). The mean distance between a nucleus' imputed position and its measured position was 11.6 spots (2.5mm). The discrepancy between imputed and measured positions was greater for cell types found throughout the embryo (e.g. skeletal muscle) than for those present in a single organ (e.g. cardiomyocytes) **(Figure 3.38F)**. Thus, despite sampling fewer transcripts than STC methods **Figure 3.38G-I)**, the retention of single cell resolution by sci-Space (that is, the ability to unambiguously ascertain sets of transcripts expressed in the same cell) affords a key advantage.

Analogously to what has been done with STC methods, but with the advantage of single cell resolution throughout, we posited that sci-Space data could serve as a scaffold for the imputation of locations of cells profiled by non-spatial single-cell RNA-seq. To test this, we aligned neurons from the developing mouse brain atlas (DMBA) to sci-Space neurons. Reassuringly, transfer of coarse DMBA anatomic annotations (from dissection) were consistent with measured sci-Space coordinates **Figure 3.40)**. Furthermore, this alignment mapped hundreds of DMBA transcriptional clusters to specific positions, many of which were spatially restricted **Figure 3.40)**. For example, of the 193 UMAP clusters from La Manno and colleagues(*103*) that mapped to slides 13 and 14,

94 and 115 clusters displayed statistically significant focal enrichment, respectively (FDR < 0.01; Getis-Ord Local G).

3.3.4 Spatial patterns of gene expression

To systematically examine these data for spatially patterned, cell type-specific gene expression across the E14.0 embryo, we quantified spatial autocorrelation – the degree to which the cells expressing a given gene are spatially proximate. Testing each annotated cell type separately, we identified hundreds to thousands of genes exhibiting positive spatial autocorrelation per cell type (Moran's test, FDR < 0.001). Amongst the cell types analyzed, connective tissue progenitors and neurons had the most spatially autocorrelated genes detected (**Figure 3.3A**; mean 12,150 $\pm$ 2,270 and 8,623 $\pm$ 3,846 genes per slide, respectively).

A "trivial" explanation for such spatial autocorrelation of genes within a cell type would be the presence of spatially-restricted, unannotated cell subtypes. For example, upon sub-clustering of connective tissue progenitors, we can indeed find spatially restricted cell subtypes **(Figure 3.42A,B)**, such that the genes defining these subtypes are naturally expected to be spatially autocorrelated. Alternatively, a gene's spatial pattern of expression could also arise from spatially restricted gene expression contributed by multiple cell subtypes. To distinguish between these scenarios, we calculated a gene's spatial autocorrelation (Spatial Moran's I) and compared this value to its autocorrelation in UMAP space (UMAP Moran's I) – a proxy for gene expression driven by a single cell subtype. Across all genes, the two measures were reasonably well-correlated (Pearson's rho 0.49; p-value $<$ 2e-16). However, a subset of genes, particularly in neurons, displayed a higher spatial autocorrelation in the tissue context than their autocorrelation in UMAP space **(Figure 3.3B; Figure 3.43A,B)**.

Closer inspection of neurons revealed that the Hox genes, a class of homeotic transcription factors that specify the body plan, were featured prominently in this subset, consistent with spatial patterning that cannot be explained solely by spatial restriction of a single cell subtype **(Figure 3.3B,C; Figure 3.43A-D)**. Genes in the HoxA cluster were expressed in a manner reminiscent

of the establishment of an anterior-posterior axis of the spinal cord (*111*) within both excitatory and inhibitory neurons **(Figure 3.3D; Figure 3.43A,B,D)**. The HoxA cluster's lack of neuronal subtype restriction was also observed in a non-spatial spinal cord dataset from the developing mouse (30) **(Figure 3.43E,F)** and validated using RNA fluorescence in situ hybridization (RNA FISH) **(Figure 3.44A-C)**.

Additional, non-Hox genes also displayed an excess of spatial autocorrelation **(Figure 3.3E,F; Figure 3.44)**. One such gene, *Cyp26b1*, is an enzyme that metabolizes the developmental morphogen retinoic acid (RA). The differential response to morphogens like RA is crucial for the patterning of tissues and leads to differential cell type specification in the developing organism (*112*). Upon examination, sci-Space data localized the focal expression of *Cyp26b1* to the brainstem, consistent with prior work (*113*, *114*), with expression observed in multiple neuronal subclusters **(Figure 3.43A,B; Figure 3.44A,B)**. This pattern of expression was observed across multiple slides **(Figure 3.44C)** and validated by RNA FISH staining of *Cyp26b1* and neuronal-subtype specific genes **(Figure 3.3G; Figure 3.44C)**. Together, these data identify *Cyp26b1* expression both in progenitors (radial glia) lining the hindbrain and in their progeny, spatially-adjacent mature neurons, suggesting that the expression of *Cyp26b1* is retained as these cells differentiate. This analysis illustrates how sci-Space can distinguish between spatial patterns of gene expression that are driven by a single cell type from those that are driven by multiple cell types.

3.3.5 Quantifying the explanatory power of spatial position

We next asked how each cell type's transcriptome varied globally across the embryo rather than on a gene-by-gene basis. We calculated the angular distance between the global transcriptomes of pairs of cells of the same type located at varying distances from one another (*115*). For many cell types, as the physical distance between cells increased, so did the angular distance between their transcriptomes. However, this trend varied by cell type, e.g. it was particularly pronounced in radial glia, neurons and endothelial cells **(Figure 3.4A; Figure 3.45A,B)**.

To quantify the contribution of spatial context to variation in gene expression across individual

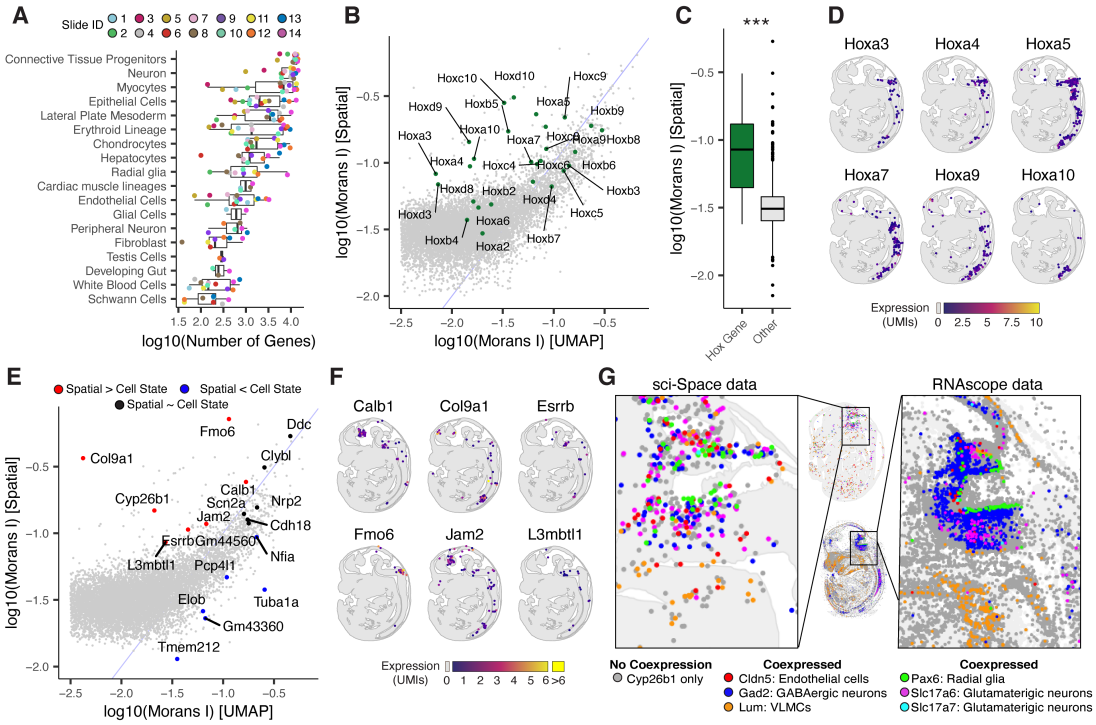

Figure 3.3: **Spatially restricted gene expression for developing neurons.**(**A**) Number of spatially significant (FDR < 0.001) autocorrelated genes within each slide (color) and cell type. Only cell types with more than 50 cells per slide were included. (**B**) Log-log (log10) plot of autocorrelation in UMAP embedding (x-axis) versus autocorrelation in spatial coordinates (y-axis) for each gene. Computed on excitatory neurons from Slide 1. Moran's I values close to 1 indicate perfect spatial correlation, while a value of 0 indicates a random spatial distribution. Hox genes are highlighted. (**C**) log10-scale boxplot of Moran's I statistic for Hox genes displayed in (**B**) versus all other expression level-matched genes (p-value < 0.001, two sided t-test). (**D**) Gene expression of HoxC cluster in Slide 1. (**E**) Similar to (**B**), log-log (log10) plot of autocorrelation in UMAP embedding (x-axis) versus autocorrelation in spatial coordinates (y-axis) for each gene with genes in different regimes highlighted for Slide 1. (**F**) Expression patterns across Slide 1 for other spatially restricted genes that are not restricted to a single neuron subcluster. (**G**) Comparison of sci-Space (shown on the more densely populated Slide 14) and RNA FISH (RNAscope) detected *Cyp26b1* patterns of expression (gray) and coexpression with markers (colors as indicated in key) for neuronal and supportive cell types.

cells, we developed a new statistical approach (Methods). Briefly, we first partitioned cells into groups based on cell type and spatial location. Then, we computed the angular distance between each cell and the average expression profile for cells of that same type in the same spatial bin. Some variance across cells is technical, due to sampling only a fraction of transcripts in each cell. We estimated the technical variance attributable to data sparsity by simulating single cell UMI count

profiles from each of the group averages. After subtracting technical variance from total variance, we were able to quantify how much of the remaining biological variance was due to each cell's type and/or spatial position. We also estimated the variance that one could expect to explain using this approach under a null model that permuted cell type and spatial position labels.

Before applying this variance decomposition approach to sci-Space data, we sought to validate it on another dataset for which sources of transcriptional variation across cells are better understood. Packer et al. previously analyzed developing C. elegans embryos, which have a defined, deterministic lineage, and observed that a cell's position in the lineage is highly predictive of its gene expression profile (*85*). We therefore decomposed variance across cells from this dataset by grouping cells that shared a common parent in the worm lineage. According to the Law of Total Variance, total variance across cells should equal the variance within these groups plus the variance between group averages. Reassuringly, this was the case **(Figure 3.46C)**. In agreement with Packer et al., our new approach showed that variance attributable to a cell's lineage-relationship in the C. elegans lineage peaked around generation 7 **(Figure 3.46D)**.

We next applied this variance decomposition approach to the sci-Space data. Variance attributable to sparse UMI sampling accounted for 95.1% of the observed variance in global gene expression across cells. In subsequent analyses, we focused on the remaining 4.9% "non-sampling" variance, and found that spatial position within the mouse embryo contributes to transcriptional heterogeneity within many cell types. Across all cells in the dataset, cell type alone accounted for 19% of the non-sampling variance in global gene expression. Moreover, a joint model that included both cell type and spatial position accounted for 50% of the non-sampling variance **(Figure 3.47)**. The implication – that spatial information explains as much if not more of gene expression variance as major cell type – was supported by the recovery of cell type and spatial gene modules of similar size and composition **(Figure 3.48)**. However, some cell types' transcriptomes appeared far more sensitive to spatial position than others (**Figure 3.4B**). For example, chondrocytes were highly influenced by position within the embryo, reflecting the ongoing development of various connective tissue lineages at E14.0, and explained at least in part by subclusters that appear to correspond to digit condensates and craniofacial mesenchyme **(Figure 3.4C; Figure 3.49)**. Other such cell types

included neurons and their precursors, the radial glia.

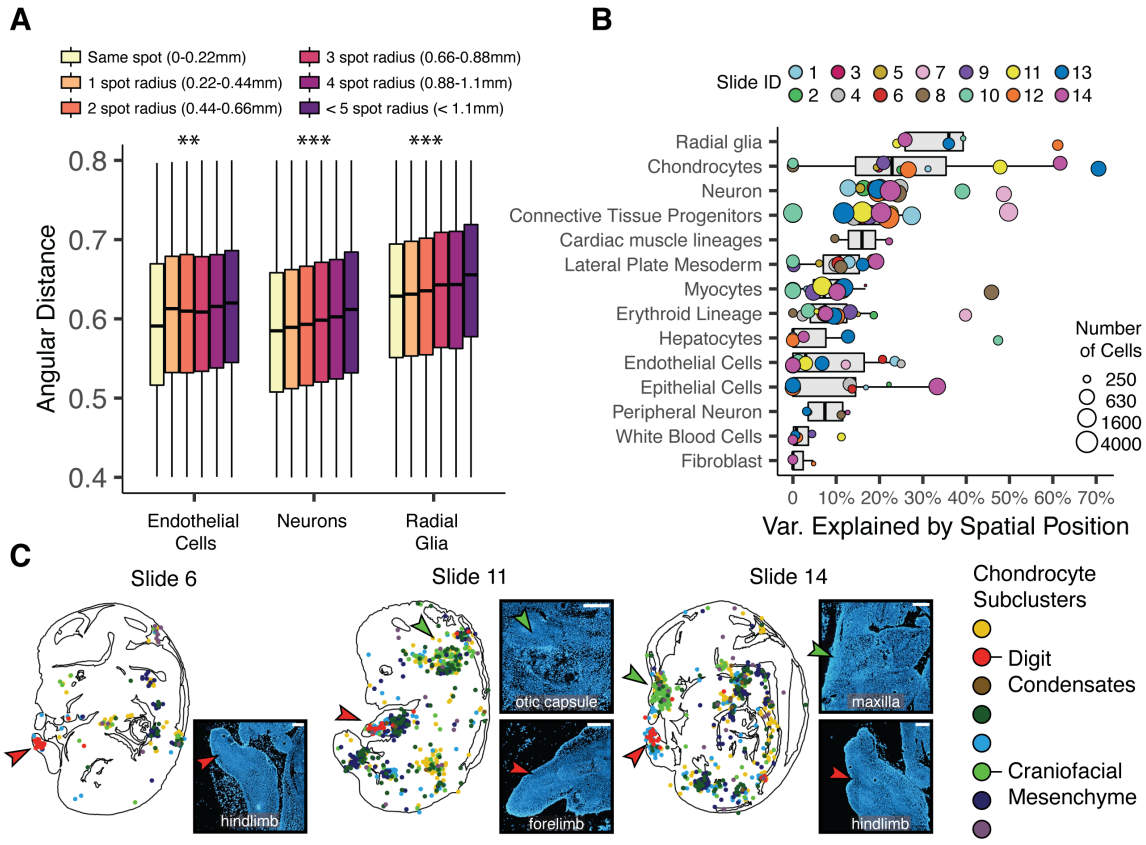

Figure 3.4: **Quantifying and characterizing the variance in gene expression attributable to spatial position.**(**A**) Pairwise angular distance (radians) between global transcriptomes of cells of the indicated cell types. Cell pairs are grouped by distance on the physical array (mm). ** : p-value < 0.001, *** p-value < 0.0001, Wald linear regression test (Methods). (**B**) Proportion of variance, apart from that attributable to sparse UMI sampling, explained within cells of each type by spatial position. Point size indicates number of cells and point color indicates the slide of origin. (**C**) Recovered positions of chondrocytes from Slides 6, 11 and 14 colored by subcluster. Arrows indicate focal concentrations of craniofacial mesenchyme (green) and digit condensate subclusters (red). Insets to the right of each plot show parts of each image with similarly positioned arrows. White text of each inset labels the anatomic structure displayed. Scale bars in panel (**C**) = 0.25 mm.

3.3.6 Pseudotemporal sci-Space trajectories reflect neuronal migration dynamics

To explore how spatial context might relate to gene expression heterogeneity in a developing cell lineage, we focused on radial glia and neurons. In particular, we hypothesized that we might be able to detect and localize the coordinated processes of neuronal differentiation and migration (*116*). UMAP dimensionality reduction of these cell types revealed the presence of three distinct trajectories originating in radial glia and leading to neurons (**Figure 3.5A**). Gene expression dynamics along these three branches were consistent with neuronal differentiation – the upregulation of cell cycle genes followed by expression of genes involved in migration **(Figure 3.50)**. Each branch was strongly enriched for subtype specific marker gene expression – *Pou4f1*+ *Pax3*+ tectal neurons, *Isl1*+/*Lhx6*+ cortical interneurons, and *Emx1*+/*Neurod6*+ cortical pyramidal neurons – indicating that the embedding captures their specification from radial glia (**Figure 3.5B**). We then examined how these trajectories were spatially distributed, by segmenting the brain from each tissue section into one of six regions using the Allen Institute's Anatomical Reference Brain Atlas (www.atlas.brain-map.org) as a guide. Cells from each trajectory overwhelmingly occupied a distinct brain region (**Figure 3.5C**). To quantify progression through differentiation, we calculated pseudotime for each cell using radial glia as the root; these values were correlated with inferred embryonic age from the developing mouse brain atlas (DMBA) (*103*) **(Figure 3.51)**. Intersecting pseudotime and spatial information, we find that cells early in differentiation clustered around the ventricles in the forebrain and developing midbrain, while those farther away exhibited a more differentiated transcriptome **(Figure 3.5E; Figure 3.51)**.

The spatial gradients of cellular maturity estimated with sci-Space data are consistent with well-documented coordination of cellular differentiation and neuronal migration. In the pallium, immature neurons migrate and differentiate radially outwards leading to the inside-out development of the cortical layers (*116*). In the sub pallium, cortical interneurons born in the ganglionic eminences migrate tangentially to populate the developing cortex and olfactory bulb (*117*). Our data also identify a third major pattern of migration in which precursors emanate from the dorsal aspect of the ventricular zone in the developing midbrain. These midbrain neurons seem to migrate

both radially, towards the pial surface, and tangentially, parallel to the pial surface, to populate this region (**Figure 3.5E** - Slides 8, 13, and 14; **Figure 3.51** - slides 4 and 7). Although radial and tangential migration are generally discussed as mutually exclusive phenomena, our data – in line with some prior work (*118*, *119*) – suggests otherwise in the developing midbrain. Furthermore, these cells share a common transcriptional program with differentiating and migrating neurons in the pallium and subpallium **(Figure 3.5F)**.

3.4 DISCUSSION

In sum, sci-Space is a new method for spatial transcriptomics that retains single cell resolution while capturing spatial information at a scale specified by a patterned array of cell hashing oligos. As a proof-of-concept, we applied sci-Space to retrieve the approximate spatial coordinates of transcriptionally profiled cells across whole mount sections from E14.0 mouse embryos. The sci-Space data are readily integrated with non-spatial single-cell RNA-seq data previously collected from mouse embryos at adjacent timepoints, enabling rapid annotation of diverse cell types and visualization of cell type-specific, spatially patterned gene expression, i.e. digital in situs. We identify examples, some expected and others novel, of genes expressed in an anatomically patterned manner within cells of a given type.

The spatial resolution of sci-Space is presently limited by the patterned array of hashing oligos, here to approximately 200 microns. Although increasing spot density and decreasing spot size are a straightforward path to increasing resolution, sci-Space is unlikely to detect effects arising from interactions between adjacent cells or to resolve gene expression gradients over short ($<$30 μm) distances. A further limitation is that we currently recover only a fraction of cells from each serial section, such that we are imputing spatially resolved maps of gene expression based on a "sample", rather than achieving dense measurements over full sections as one would obtain from in situ hybridizations, spatial transcriptomics or Slide-seq.

Nonetheless, sci-Space fills a key need not addressed by these or other technologies. Like other spatial transcriptomic methods relying on patterned arrays (e.g. Slide-seq), sci-Space can be

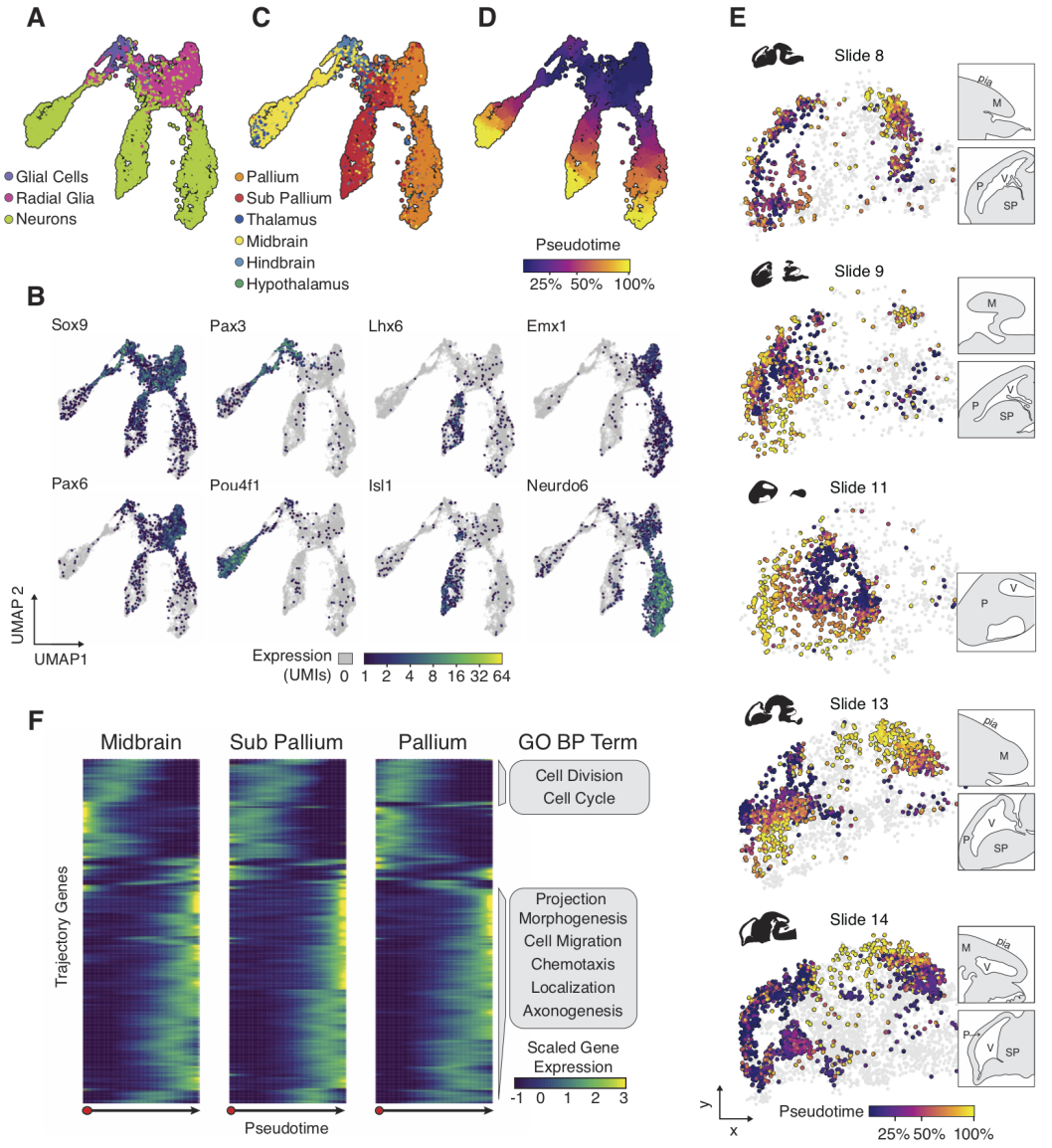

Figure 3.5: **Pseudotemporal spatial trajectories capture migratory patterns in the developing brain.** (**A-C**) UMAP embedding displaying the neural trajectories colored by (**A**) cell type, (**B**) specific gene expression, (**C**) cortical region, or (**D**) pseudotime. (**E**) Spatial positions of neurons and radial glia in the cortex colored by pseudotime if a neuron was found in the trajectory or otherwise colored grey. Insets of rostral (below) and caudal (above) brain outlines are shown to the right of each slide (M-Midbrain, P-Pallium, SP-Sub Pallium, V-Ventricle). (**F**) Scaled and centered gene expression for genes (rows) significantly varying over pseudotime in all three trajectories. Enriched Gene Ontology Biological Processes terms (GO BP) are displayed next to clustered genes.

applied, routinely and efficiently to large regions, e.g. whole embryo serial sections. However, unlike these methods, sci-Space recovers single cell transcriptomes. It can therefore capture patterns of spatial gene regulation private to specific cell types and estimate the contribution of each cell type to the expression of morphogens and other signalling molecules, both within and across anatomical regions. Moreover, sci-Space data can serve as a spatial "scaffold" for conventional, non-spatial single-cell RNA-seq atlases, which may be considerably more challenging to map onto tissues profiled by spatial profiling methods that lack single-cell resolution.

At the other end of the methodological spectrum, seqFISH, MERFISH, FISSEQ and similar multiplex mRNA imaging methods do provide single cell resolution and are readily capable of detecting gene expression gradients consequent to cell-cell interactions. On the other hand, the complex instrumentation and long imaging times that these technologies require are impractical to routinely apply at the scale of entire tissues, organs or embryos. Furthermore, at least for seqFISH and MERFISH, genes are profiled through targeted hybridization with gene-specific probes that must be designed and validated. In contrast, the imaging steps of sci-Space can be performed on a conventional widefield microscope in a few minutes per slide, and capture the global gene expression program via untargeted 3' end mRNA sequencing.

Finally, we developed a statistical approach to identify cell types in the developing embryo that exhibit spatially regulated gene expression. This method quantified the contributions of spatial position, cell type, and technical factors to decompose the variance across cells' transcriptomes. We found that in some cases, spatial context explained as much or more variance as cell type. Closer analysis of radial glia and neurons revealed gradients of developmental maturity in different regions of the brain indicative of known and novel patterns of neuronal migration. Together with data from complementary technologies, we anticipate that the further application of sci-Space to serial sections spanning entire embryos from many timepoints will facilitate the construction of a set of highly time- and space-resolved 4-dimensional atlases of gene expression across the entirety of mammalian development.

3.5 ACKNOWLEDGMENTS

We thank members of the Lampe, Stevens, Shendure, Trapnell lab and others for their critical feedback, particularly Eva Nichols, Riza Daza, Junyue Cao, Vijay Ramani, and Lauren Saunders for helpful discussions during the development of the protocol and analysis of the results. We also thank Choli Lee and Dana Jackson for supporting the research environment. Funding: Aspects of this work were supported by funding from the NIH (1R01HG010632 to C.T. and J.S.; DP2HL137188 to K.R.S.), an NIH/NIBIB training grant (T32EB1650 to S.S.), the Brotman Baty Institute for Precision Medicine, the Paul G. Allen Frontiers Foundation (Allen Discovery Center grant to J.S. and C.T.), the Chan Zuckerberg Initiative (to C.T.), the Washington Research Foundation Postdoctoral Fellowship (to M.C.R.). J.S. is an investigator of the Howard Hughes Medical Institute. Author contributions: S.R.S. and M.C.R. developed sci-Space and performed the experiments. J.L. aided in the fabrication of sci-Space slides with the support of P.L.; S.R.S. and M.C.R. performed the computational analyses with assistance from E.B., P.G., J.F., M.D. and J.P. and under the supervision of C.T.; Immunostaining and RNA FISH experiments were performed by M.C.R. with assistance from S.S.; M.S. and C.L. provided critical feedback on the project; S.R.S., M.C.R., C.T., K.R.S. and J.S. wrote the manuscript with input from all coauthors; and K.R.S. J.S. and C.T. supervised the project. Competing interests: One or more embodiments of one or more patents and patent applications filed by the University of Washington may encompass the methods, reagents, and data disclosed in this manuscript. Data and materials availability: Processed and raw data can be downloaded from NCBI at GSE166692. All the code used to perform the presented analyses is hosted on github and indexed on Zenodo. Developed laboratory protocols can be found on protocols.io (see Methods).

3.6 METHODS

3.6.1 Overview

Sci-Space is a single cell technology wherein positional information is recorded and read out in conjunction with single cell transcriptomes. This process is mediated by the spatial transfer of hashing oligos (which resemble poly-adenylated (polyA) transcripts) onto cells. A combination of 3 hashing oligos is meant to uniquely specify the slide from which a cell originates as well as its position within a regular grid on that slide. After performing sci-RNA-seq library preparation and sequencing, the spatial barcodes found within each cell's library are used to assign it to a slide as well as to position it within the grid. The molecular biology details needed to perform the sci-Space procedure and library preparation are outlined below as well as in a series of published protocols.io sites:

- Creating sci-Space grids for spatial barcoding - https://www.protocols.io/view/creating-sci-space-grids-for-spatial-barcoding-bm64k9gw
- Spatial transfer of oligonucleotides and imaging - https://www.protocols.io/view/spatial-transfer-of-oligonucleotides-and-imaging-bqt7mwrn
- Single cell RNA sequencing library preparation - https://www.protocols.io/view/single-cell-rna-sequencing-library-preparation-2-l-bquamwse

3.6.2 Creating sci-Space grids for spatial barcoding

A thin membrane of dried agarose was fabricated on the surface of microscope slides (Superfrost Plus, Thermofisher). This agarose matrix absorbed and retained an array of spotted oligo hashes. To prepare nuclease-free agarose, 3% w/v low melting temperature agarose powder (SeaPlaque, Lonza, Bend, OR) was added to deionized water containing 0.1% v/v diethyl pyrocarbonate, incubated 2 hr at room temperature, and autoclaved for 15 min. The uniform thickness of the layer of agarose across the slide surface was patterned using spacers of two stacked 22 x 22 mm, number

one thickness (0.15 $\pm$ 0.02 mm each) coverslips overhanging either end of the slide. Molding of the agarose was performed by pipetting a 300 μL volume of heated agarose solution into the center of the slide and slowly placing a second slide onto the agarose solution avoiding the formation of bubbles. The molding slide was allowed to rest on the cover glass spacers. After the agarose had gelled between the two slides (30-60 min on ice) a razor blade was used to release the exposed edges of the agarose layer from the top, molding slide. The two slides were then carefully slid apart and the cover glass spacers were removed. The resulting thin layer of agarose gel was dried onto the bottom slide overnight in a biosafety cabinet. All agarose slides were UV-treated for 20-30 min prior to spotting to further protect against nuclease activity.

The space-grid array of hashing oligos and SYBR green reference points was spotted onto agarose-coated slides using a QArray2 microarray scanner (Genetix, New Milton, Hampshire, GB). A series of 384-well high sample recovery plates (Molecular Devices, San Jose, CA, X7020) was prepared containing a final concentration of 15μM spot oligo and 2.5μM sector oligo per well (Integrated DNA Technologies, Coralville, IA), and 0.5% v/v glycerol, with or without SYBR green dye ([5x] Thermofisher, S7585) to achieve the predetermined oligo and SYBR green reference point layout when a 21 x 21 spot/pin array was printed with 16 spotting pins (4 x 4 grid). These printing parameters gave space-grids containing 7056 (84 x 84) spots of unique oligo combinations. The spotting height was adjusted to ensure consistent contact of the spotting pins with the transfer slides' agarose coating.

3.6.3 Testing blotting concentrations of hash oligos and SYBR green

Space-grids for testing hash oligo blotting concentrations were prepared as noted above using the QArray2 microarray scanner (Genetix, New Milton, Hampshire, GB). Each space-grid was given a single distinct DNA barcode sequence at a chosen final concentration (10μM, 20μM, 25μM or 50μM) with a single sector marked with 5x SYBR green. These space-grids were then blotted onto a series of mouse embryo sections ranging from E13 to E16 (Zyagen, San Diego, CA), by sandwiching the two slides (tissue and space-grid). First, permeabilization and hashing solution

was prepared for each slide by mixing a unique slide-specific hash oligo (5μL at 10μM) in the 495μL permeabilization solution [10mM Tris/HCl pH 7.4, 10mM NaCl, 3mM MgCl2 with 1% v/v superase inhibitor (Invitrogen) and 0.1%v/v IGEPAL CA-630 (Sigma Aldrich)]. Following permeabilization, each slide was barcoded via transfer with a test space-grid. The transfer was then imaged and the cells were harvested by cell scraping into a solution of 5% paraformaldehyde (cat no. 100504-940, VWR) in 1x PBS. This is described in detail in a subsequent section ("Spatial Transfer of Oligonucleotides and Imaging"). After 15 minutes of fixation on ice, cells were centrifuged (800g for 10 minutes), pooled and subjected to sci-RNA-seq2 library preparation (*120*). This is described in detail in a subsequent section ("Single cell RNA sequencing library preparation").

To test co-cultured human and mouse cells, HEK293T cells and NIH3T3 cells were placed in a droplet on a coverslip coated with 1% gelatin. 4 coverslips were prepared and cells were allowed to attach overnight to the coverslip surface. Next, each coverslip was permeabilized as detailed above and labeled with a gel containing a single hash oligo. These nuclei were then scraped into paraformaldehyde, fixed for 15 minutes on ice and subject to sci-RNA-seq library preparation as detailed below.

3.6.4 Testing spotted space-grids

Hash oligos from three space-grids were dissolved in 500μL of permeabilization solution [10mM Tris/HCl pH 7.4, 10mM NaCl, 3mM MgCl2 with 1% v/v superase inhibitor (Invitrogen) and 0.1%v/v IGEPAL CA-630 (Sigma Aldrich)]. Concurrently, three aliquots of 2 million HEK29T cells were harvested and washed once with 1x PBS. The resuspended hash oligo solutions were then used to lyse and label the HEK293T nuclei. After a 3 minute incubation on ice, the nuclei suspension was chemically fixed with 5mL of 4% paraformaldehyde and incubated for 15 minutes on ice. Nuclei were then pelleted at 500g for 5 minutes, washed, with 500μL Nuclei Buffer (NSB) [10mM Tris/HCl pH 7.4, 10mM NaCl, 3mM MgCl2 with 1% v/v superase inhibitor (Invitrogen) 1% v/v BSA (New England Biolabs)] and permeabilized by resuspension in 500μL NB + 0.2%

Triton-X. These nuclei were centrifuged and washed with 500uL of NSB. These nuclei were then pelleted and 5000 nuclei from each sample was loaded into indexed reverse-transcription reactions. Reverse transcription was performed as described previously (*120*) were pooled and 25 nuclei were sorted into a 96 well plate containing 16μL of elution buffer per well. Libraries were prepared by performing an indexed PCR using 20uL of NEBNext High-Fidelity 2X PCR Master Mix (NEB), 2μL of 10μM indexed P5 primer and 2μL of 10μM indexed P7 primer. PCR was run for 18 cycles with the following settings: 72°C for 5 min, 98°C for 30 sec, 18 cycles of (98°C for 10 sec, 66°C for 30 sec, 72°C for 30 sec) and a final 72°C for 5 min. These libraries were then pooled and sequenced on a Nextseq 500 (Illumina, San Diego, CA) using a high output 75 cycle kit (Read 1: 18 cycles, Read 2: 52 cycles, Index 1: 10 cycles and Index 2: 10 cycles).

3.6.5 Spatial transfer of oligonucleotides and imaging

Serial sections of an E14.0 mouse embryo were purchased (Zyagen, San Diego, CA) and stored at -80°C prior to use. Oligo hashes were transferred in their arrayed pattern from the space-grid slides to fresh-frozen embryo sections by diffusion through cell permeabilization buffer. First, the embryo slide was placed so that it rested (tissue facing up) with the tissue section between two transfer clips. Subsequently, 500μL of nuclei permeabilization buffer [10mM Tris/HCl pH 7.4, 10mM NaCl, 3mM MgCl2 with 1% v/v superase inhibitor (Invitrogen) and 0.1%v/v IGEPAL CA-630 (Sigma Aldrich)] with 5μL of slide-specific hashing oligo at 10μM and 5μL of 500μM stock DAPI, was pipetted gently onto the tissue section and across the long edge of the embryo slide nearest the user using an wide bore p1000 tip. A space-grid transfer slide was then positioned (agarose surface facing the tissue section) so that the arrayed oligos were aligned between the two transfer clips and spanned the tissue section's extent. Placement of the space-grid slide was achieved by tilting the slide so that its long edge nearest the user contacted the edge of the tissue section slide and fit under the overhanging fastening teeth of the transfer clips. The space-grid slide was then rocked toward the embryo section slide until the two slides were face-to-face with the tissue section contacting the space-grid oligo array-laden agarose membrane. Excess buffer

was allowed to wick into a laboratory wipe. When stacked, the slides snapped into the transfer clips and were thereby securely held together during transfer. The slide stack was moved to the microscope stage and the entire embryo section was imaged in GFP and DAPI channels. The transfer slide was then removed from the transfer clips and separated from the tissue.

Cells of the embryo section were then scraped using a cell scraper (Fisherbrand, GDPC240) from the slide into a 4% paraformaldehyde fixing solution. At this point, Slides 7-14 were subjected to sonication using the bioruptor sonicator (Diagenode, product number B01020001). This extra round of sonication aided in disassociating more nuclei. After fixation for 15 minutes on ice, cells were spun down in 1.5mL tubes in a chilled benchtop centrifuge at 800g for 10 minutes. The supernatant in each tube was removed and cells were pooled in 1mL of NSB [Nuclei Buffer (10mM Tris/HCl pH 7.4, 10mM NaCl, 3mM MgCl2) with 1% v/v superase inhibitor (Invitrogen) and 1% v/v BSA (New England Biolabs)], flash frozen and stored at -80C.

3.6.6 Single cell RNA sequencing library preparation

Frozen nuclei were thawed over ice and spun down at 800g for 8 minutes in a 15mL conical tube. Cells were then permeabilized in 400μL of permeabilization buffer (NSB + 0.25% Triton-X) for 3 minutes and then spun down at 800g for 8 minutes. Following resuspension in 500μL NSB, two-level sci-RNA-seq libraries were prepared as previously described (*98*, *120*). Briefly, nuclei were first sonicated for 12 seconds using the bioruptor sonicator on the low setting. This caused disruption of many nucleus aggregates that had formed. Cell counts were then obtained by staining nuclei with 0.4 % trypan blue (Sigma-Aldrich) and counted using a hemocytometer.

5000 nuclei in 2μL of NSB and 0.25μL of 10mM dNTP mix (Thermo Fisher Scientific, R0193) were then distributed onto a skirted twin.tec 96 well LoBind plate (Fisher Scientific, 0030129512). Next, 1μL of uniquely indexed oligo-dT at 25μM (*120*) was added to every well and mixed. This 96-well plate was then incubated at 55C$^\circ$ for 5 minutes and then immediately placed on ice. Next, 1.75μL of reverse transcription mix (1μL of Superscript IV first-strand buffer, 0.25μL of 100 mM DTT, 0.25μL of Superscript IV and 0.25μL of RNAseOUT recombinant ribonuclease inhibitor)

was then added to every well and mixed. Plates containing the reverse transcription reactions were then incubated using a ramping reverse transcription protocol on a thermocycler:

1. 10C° for 2 minutes
2. 20C° for 2 minutes
3. 30C° for 2 minutes
4. 40C° for 2 minutes
5. 50C° for 2 minutes
6. 55C° for 15 minutes
7. 4C° forever

Wells were pooled and nuclei were transferred to a flow cytometry tube through a 0.35 μm filter cap and DAPI added to a final concentration of 3 μM. At this point pooled nuclei were seeded using one of 2 methods; (1) sorted on a BD FACS Aria II cell sorter (Slides 1-6) or (2) diluted (Slides 7-14).

Sorted cells from Slides 1-6 were prepared using 8 RT plates. After calculating the collision rate using the birthday problem calculator (*121*), we sorted 200 nuclei per well into 96-well LoBind plates with each well containing 5μL of EB buffer (Qiagen) and 0.75μL of second strand mix (0.5 μL of mRNA second strand synthesis buffer and 0.25μL of mRNA second strand synthesis enzyme, New England Biolabs).

Diluted nuclei originating from slides 7-14 were prepared using 10 RT plates were first diluted to 50 nuclei per μL in NSB. Diluted nuclei were then premixed with second strand-synthesis reagents (0.5 μL of mRNA second strand synthesis buffer per 5μL of nuclei suspension and 0.25μL of mRNA second strand synthesis enzyme per 5μL of nuclei suspension). 5.75μL of

this nucleus/second-strand synthesis solution was dispensed into 96-well LoBind plates to seed 250 nuclei per well.

Second strand synthesis performed at 16C° for 150 minutes. Tagmentation was then performed by addition of 5.75μL of tagmentation mix per well (0.01μL of a custom n7-loaded Tn5 enzyme in 5.74μL 2x Nextera TD buffer, Illumina) and plates incubated for 5 minutes at 55C°. This reaction was terminated by addition of 12μL of DNA binding buffer (Zymo, D4004-1-L) and incubated for 5 minutes at room temperature. 36μL of Ampure XP beads (Beckman Coulter, A63880) were added to every well, DNA purified using the standard Ampure XP protocol eluting with 17μL of EB buffer. 16μL of this eluate was then transferred to a new 96 well LoBind plate for index PCR.

For PCR, 2μL of 10μM indexed P5, 2μL of 10μM indexed P7 (*120*) and 20μL of NEBNext High-Fidelity master mix (New England Biolabs, M0541L) were added to 16 uL of eluted template DNA. PCR indices were arrayed such that each well contained a unique combination of P5 and P7 barcodes. PCR was then performed using the following program:

1. 75C° for 3 minutes
2. 98C° for 30 seconds
3. 98C° for 10 seconds,
4. 66C° for 30 seconds
5. 72C° for 1 minute
6. Return to Step 3 for 17 times
7. 72C° for 5 minutes.
8. 4C° forever

After PCR, all wells were pooled, concentrated using a DNA clean and concentrator kit (Zymo Research, D4033) and purified via a 0.8X Ampure XP cleanup. Final library concentrations were determined by Qubit (Invitrogen), libraries visualized using a TapeStation D1000 DNA Screen tape (Agilent, 5067-5582) and libraries sequenced on a Nextseq 500 using a high output 75 cycle kit (Illumina, 20024906). Libraries were denatured and diluted to 2pM. Sequencing was performed using onboard primers using the following sequencing recipe:

- Read 1: 18 cycles (8 base-pair UMI and 10 base-pair RT barcode
- Index 1: 10 cycles (10 base-pair PCR index1)
- Index 2: 10 cycles (10 base-pair PCR index2)
- Read 2: 52 cycles (52 bases of transcript or hash-oligo)

3.6.7 Pre-processing of sequencing data

Sequencing data was processed as described previously (*98*). Briefly, sequencing runs were first demultiplexed using bcl2fastq v.2.19. Only barcodes that matched reverse transcription indices within an edit distance of 2 bp were retained. Following assignment of indices, polyA tails were trimmed using trim-galore (https://github.com/FelixKrueger/TrimGalore), and reads were mapped to a mouse transcriptome (mm-10), human-mouse transcriptome (hg-38 and mm-10) using the STAR aligner. Following alignment, reads were filtered for alignment quality, and duplicates were removed. Reads were considered duplicates if they (1) mapped to the same gene, (2) mapped to the same cell barcode and (3) contained the same unique molecular identifier (UMI). Reads that met the first two criteria, and differed by an edit distance of 1 from a previously observed UMI were also marked as duplicates and discarded. Non-duplicate reads were assigned to genes using bedtools to intersect with an annotated gene model. All 3' UTRs in the gene model were extended by 100 bp to account for the possibility that some gene 3' UTR annotations may be too short, causing genic reads to improperly be annotated as intergenic. Cell barcodes were considered

to correspond to a bona fide cell if the number of unique reads associated with the barcode was greater than an interactively defined threshold on a knee plot. Reads from cells that passed this UMI count threshold were first aggregated into a sparse matrix format and then loaded and saved as a CDS object for analysis with Monocle3 or Seurat.

3.6.8 Slide registration

To map cell locations to an image of the embryo, fluorescent SYBR green spots with known positions are used to orient cells. Images of the hash array with fluorescent SYBR green spots on top of the DAPI-stained embryo section were taken with a 2.5x magnification (Zeiss Observer Z1 Microscope).

Prior to the transfer of the image of Slide 12, the Slide 12's image file was accidentally erased and could not be recovered. The sequenced cells mapping to Slide 12 are included in the dataset with relative spatial positions based on the recovered hash oligos. However, for this slide we were unable to perform any analyses that relied on segmentation or slide registration.

The captured images were then used to orient the hash array to the embryo section. More specifically, co-registration of the imaged embryo sections and the oligo hash tagged transcriptomes was achieved through alignment of the SYBR green waypoints imaged during transfer to their position within an ideal space-grid layout. Coordinates for SYBR green spots imaged during oligo hash transfer to the embryo section and the corresponding coordinates in an image of an ideal space-grid were obtained in Fiji image processing software (*106*) using the Big Warp function of the BigDataViewer plugin (*107*). An affine matrix was computed using the coordinates as source (embryo image) and target (space-grid image) control points in the AffineTransformation function in the "vec2dtransf" and "imager" packages in R. The matrix was applied to the embryo section image. Sequenced nuclei were then mapped to the aligned space of the transformed image and space-grid using their space-grid hashes.

The following formula was used to calculate the number of microns per pixel and thereby estimate the size of each spot on the hash array.

microns per 1 pixel = (native camera pixel size / objective / camera adaptor)

One pixel was equal to 1.816 microns based on the camera pixel size of 4.54 for a Zeiss Axiocam 503 Mono Camera, an objective size of 2.5x and a camera adaptor size of 1.

3.6.9 Assigning spatial labels from hash reads

Reads from hash oligos were demultiplexed as described previously (*98*). Briefly, demultiplexed reads that matched combinatorial indexing barcodes were examined to identify hash reads. Reads were considered hash reads when they met two criteria: 1) the first 10 bp of read 2 matched a hash barcode in the experiment within an edit distance of two; and 2) contained a polyA track between base pairs 12 to 16 of read 2. These reads were then deduplicated by cell barcode and collapsed by UMIs to create a vector of hash oligo UMI counts for each nucleus in the experiment.

To assign each nucleus to the slide from which it came, we tested whether its sci-RNA-seq library was enriched for a particular hash barcode. We compared a nucleus's hash UMIs against a 'background distribution', which under ideal circumstances, would be random and uniformly distributed. To estimate the background distribution, we simply aggregated the hash UMIs from cell indices for which fewer than 10 mRNA UMIs were collected, reasoning that these reflect library contributions from ambient reverse transcriptase products, debris fragments, etc. We then compared the hash UMIs for nucleus to this background by a chi-squared test. After correcting the resulting p-values were corrected for multiple testing by the Benjamini-Hochberg procedure, we rejected the null hypothesis that originates from the background distribution at the specified FDR (5% FDR was used in this study). Those nuclei with hash counts deemed greater than background were also evaluated for enrichment for a single hash sequence. Enrichment ratios were calculated as the UMI count ratio of the most abundant vs. the second most abundant hash oligo. Specifically, if the UMI count for the most abundant hash in a nucleus is α-fold higher than the second most abundant, is marked as a singleton. α was set to 5, which corresponded to a nadir between two modal outcomes – separated unlabeled cells and singularly labeled cells. Cells that fell below 5-fold enrichment of a unique hash oligo were flagged as a multiplet or debris and discarded.

A cell's spatial position within the grid consists of a specific combination of two oligonucleotides, a spot oligo and a sector oligo. To find a cell's position within the grid, we first took a single cell's vector of spot counts and mapped these counts to their position in the sci-Space grid. Then we performed a 3x3 gaussian convolution on every position, allowing us to account for and integrate spatially local signal from neighboring positions and simultaneously reduce spurious background. Next, we performed an element-wise multiplication of convolved spot values by the matrix of sector counts measured for that cell. The product of spot and sector oligos were then ranked and a cell was mapped to the top ranking combination which matched two criteria: (1) the combination represented a valid pairing and (2) the combination mapped within the boundary of the imaged embryo. This boundary was determined by manually segmenting the outline of the DAPI stained image of the embryo.

Upon manual inspection of these draft spatial positions, we noticed a minority of nuclei which were clearly mismapped (e.g. cardiomyocytes outside of the heart). This seemed to occur through a nucleus' absorption of a neighboring sector oligo or alternatively via barcode collision between cellular debris carrying a spatial index and the sequenced nucleus. To correct these misassignments, each slide was manually segmented into regions of interest (ROI) where a cell type was focally concentrated. Next we asked whether a nucleus of a given cell type had a plausible alternate mapping within these ROIs. A nucleus was moved from its draft position if the nucleus' highest convolved spot value within the region of interest was within 5-fold of its highest spot value. The value of 5-fold was chosen by examining the distribution of all alternate mappings. This procedure was then repeated for each cluster using an automated algorithm to detect high density regions of clusters within a slide. Through this process, the spatial calls of 9.1% nuclei were remapped. Finally, we removed cardiomyocytes and hepatocytes that mapped outside of the heart and the liver, respectively, resulting in the removal of 369 nuclei.

3.6.10 Estimating nuclei counts from embryo images

Using python, the 2.5x magnification embryo DAPI stained images were preprocessed using a white top-hat transform followed by a histogram equalization to reduce uneven lighting and increase contrast respectively. The images were then thresholded using Otsu's method. In order to overcome the challenge of counting individual cells in nuclei clusters, the resulting binary masks were separated into 'dense' and 'sparse' nuclei masks using a connected components algorithm. The dense nuclei masks were used to isolate nuclei clusters in the original embryo images, which were then thresholded to a secondary value defined as Otsu's value plus a constant intensity shift. The sparse and the dense nuclei masks were then distance transformed, using Euclidean geometry, to generate distance maps. A peak finding algorithm was used to isolate the centroids of peaks in the distance maps and resulting unique centroids were counted as nuclei.

3.6.11 Cell type classification

Nearest neighbor classification was performed by aligning the cells from this study to cells from E13.5 time point from the MOCA single cell dataset (*84*). This MOCA dataset was chosen because it was prepared using nuclear sci-RNA-seq. It is our experience that alignment between datasets produced using the same technology are less sensitive to hyperparameter selection during alignment. The E13.5 time point was chosen because this time point most closely matched the E14.0 timepoint sequenced in this study.

Count matrices from the two datasets were subsetted for genes found in both datasets and then combined. The E13.5 time point was then downsampled and the two datasets were aligned using Seurat v3(*102*) dataset integration using reciprocal PCA. For each cell in the E14.0 time point the 10 nearest neighbors in UMAP space from the MOCA dataset were recorded. Each cell was then assigned the majority nearest neighbor label. Finally, to remove poor confidence cell type labels, the E14.0 data was clustered and cell type labels that did not account for more than 5% of a cluster were assigned "Unknown". Garnett classification (*104*) was performed using a marker-free classifier trained on the E13.5 time point from the MOCA dataset. This classifier was then applied

to the cells sequenced in this study.

This same process was repeated to match neurons from the developing mouse brain atlas dataset (DMBA) to neurons in the sci-Space dataset(*103*). Briefly, neurons and radial glia from the E13.5, E14.0 and E14.5 timepoints were used to perform nearest neighbor alignment using reciprocal PCA (*102*). The majority label from a cell's 5 nearest neighbors was used to transfer a number of different labels including cell type, anatomical dissection, age, and UMAP cluster in the original dataset. Finally, the different inferred labels were collated along with the top differential genes marking each cluster. The final annotation set consists of a combination of cell types transferred from the MOCA and DMBA datasets along with manual annotation matching a cluster's differential gene expression.

3.6.12 Immunostaining and adjacent image alignment

Before immunostaining serial sections adjacent to sequenced sections were fixed in 4% paraformaldehyde (Electron Microscopy Sciences) in phosphate buffered saline (PBS, ThermoFisher) for three minutes at room temperature. Sections were then washed for five minutes three times in PBS-T (0.1% Tween-20, VWR), permeabilized for 10 minutes at room temperature in 0.1% Triton X-100 (VWR) in PBS, and washed for five minutes three times in PBS. Autofluorescence was quenched using TrueBlack Lipofuscin (Biotium) according to the manufacturer's protocol. Briefly, sections were treated with TrueBlack Lipofuscin diluted 20X in 70% ethanol with a 30 second to three minute incubation, then washed for five minutes three times with PBS. Sections were next blocked with 2.5% normal donkey serum (NDS, Jackson ImmunoResearch Laboratories) in PBS for one hour at room temperature. Primary antibodies were applied in PBS containing 2.5% NDS as indicated in Table S1 with an overnight incubation at 4ºC. The following day sections were washed for five minutes three times in 2.5% NDS in PBS, then incubated for one hour at room temperature with Hoechst 33342, Trihydrochloride, Trihydrate (Invitrogen) counterstain and secondary antibodies diluted as indicated in Table S1 in 2.5% NDS in PBS. Sections were next washed again in 2.5% NDS in PBS and then coverslipped with Fluoromount-G Mounting Medium (Southern-

Biotech) prior to imaging.

Stained sections were imaged using a Ti-E inverted microscope (Nikon) and multi-field images were stitched in the NIS-Elements (Nikon) software. Each channel of the triple stained sections together with the counterstain was aligned to the DAPI image of the sequenced section using the StackReg plugin in Fiji (rigid body followed by affine). Each aligned channel of the stained section images were then separated from the counterstain and overlaid onto the DAPI channel image of the adjacent sequenced section.

3.6.13 Anatomical annotation and segmentation

Annotation was performed using The Atlas of Mouse Development (*122*) in conjunction with magnetic resonance images of the E14.5 embryo (*123*). Annotations were then confirmed using immunostained adjacent sections (when available). Anatomical segmentation was performed manually using the DAPI-stained embryo section and the Big Warp function of the BigDataViewer plugin in Fiji. The region of interest was demarcated by choosing a bounding set of points in clockwise or counter-clockwise order. These points were then used to construct a polygon using the spatial features package in R. These polygons were then scaled using the same affine transformation used for slide registration to put them on the same coordinate axis. For polygons with holes, the contour of the entire image was first segmented followed by segmentation of each cavity. This same process was repeated to segment and annotate the brain regions. However, the Allen Institute's Anatomical Reference Brain Atlas (www.atlas.brain-map.org) was used as a guide to annotate the Pallium, Sub Pallium, Midbrain, Hindbrain, Thalamus and Hypothalamus of each brain.

3.6.14 Tissue domains

Tissue domains with similar cell type composition were identified using Giotto (*108*), an unsupervised tool for single-cell spatial expression analysis. The top 500 spatially autocorrelated genes (Moran's I > 0.05, FDR < 0.001, expressed in at least 1% of cells) were identified from all slides.

For each slide, spatial domains were identified using the Hidden Markov random field model with parameters k=50, beta=10. To identify consensus spatial domains that existed on multiple slides, UMAP dimensionality reduction was performed based on the absolute numbers of annotated cell-types in each domain from each slide. Community detection with louvain clustering (k=5) identified 22 clusters representing the 22 tissue domains.

3.6.15 Analysis of aggregated spatial positions

Nuclei mapping to the same spatial grid position within each slide were aggregated by summing gene expression counts from each nucleus. These spatial positions were then treated as the columns of a gene by position count matrix. This matrix was converted into a Monocle3 CDS object, and used as input for PCA, UMAP and louvain clustering.

To assess the ability of integration to recover a cell's spatial position, we used the Seurat package (version 3.2.2) and applied this analysis to Slide 14. First, nuclei mapping to a spatial position were aggregated and used to create a Seurat object bearing the spatial coordinates of each position. These spatial positions were then processed as described by the Seurat spatial vignette. Briefly, this involved running SCTransform, PCA, UMAP and clustering using the default parameters. The nuclei used to aggregate each of these positions (nuclei from Slide 14) were also processed in the same way (SCTransform, PCA, UMAP and clustering). To integrate the two datasets we used the FindTransferAnchors() function in Seurat to find anchors between the two datasets using the SCTransform as the normalization method. Finally, these anchors were used as an input to the TransferData() function to return a set of predicted spatial positions for each cell. To determine the error associated between data integration and the pre-aggregation ground truth, a cell was assigned to the position with the highest transfer probability returned by TransferData() and the euclidean distance was calculated between the most probable transferred position and the ground truth position.

3.6.16 Kriging gene expression

Spatial grid positions were first collapsed such that non-overlapping sets of 4 adjacent positions were collapsed into a single spatial position. Each cell type within these spatial bins was then aggregated by summing the counts for each gene contributed by that cell type. These values were then kriged with the automap package in R using ordinary kriging via the autoKrige() function. Interpolated values were then rescaled to reflect the percentage of gene expression contributed by a cell type at each given position. Finally, a polygon object specific to each slide was used to clip the interpolated gene expression values.

3.6.17 Spatial autocorrelation analysis

Gene spatial autocorrelation was computed by first subsetting cell types for which there were more than 100 cells present on a slide. After setting a random seed, we estimated size factors, performed PCA and UMAP dimensionality reduction with a fixed set of parameters. UMAP was run using uwot's implementation in R with the flag fast_sgd set to false. This ensured that UMAP dimensionality reductions were consistent between runs. Following UMAP, for each subset of cells, a gene's spatial autocorrelation was computed using either its cell's spatial coordinates or UMAP coordinates as the input into Monocle3's graph_test() function. The resulting test statistic was corrected for multiple testing and genes with an FDR < 0.01 and a Moran's I test statistic greater than 0.05 were reported as having statistically significant spatial autocorrelation.

3.6.18 RNA Fluorescence in situ Hybridization (FISH) and analysis

To validate spatial expression patterns across cell types identified in our dataset, an 8-probe RNAscope HiPlex kit (Advanced Cell Diagnostics, Inc.) including probes against 6 transcripts to mark various cell populations: *Gad2* – GABAergic neurons; *Slc17a7* – VGlut1+ glutaminergic neurons; *Slc17a6* – VGlut2+ glutaminergic neurons; Pax6 – radial glia; Lum – fibroblasts; and finally, *Cldn5* – endothelial cells. The kit additionally assayed for *Cyp26b1* and *Hoxa10* to demonstrate their spatial localization within multiple cell types and neuronal cell subtypes as indicated by our

sciSpace data. Briefly, serial sections near Slide 1 of the sci-Space dataset were assayed according to the manufacturer's protocol. The fresh-frozen tissue sections were fixed using 4% paraformaldehyde (Electron Microscopy Sciences) in 1X PBS, dehydrated, and treated with the Protease IV kit component. Following the manufacturer's specified hybridization steps, counterstaining, and coverslipping, the first four probes: 1) *Cyp26b1*, 2) *Gad2*, 3) *Slc17a6*, 4) *Hoxa10* were imaged. Scans were obtained using an Aperio VERSA slide scanner (Leica Biosystems) with 40x magnification and DAPI, FITC, Cy3, Cy5, and Cy7 filter sets. Coverslips were removed, the first four fluorophores were cleaved, the fluorophores for probes 5-8 were hybridized: 5) *Pax6*, 6) *Slc17a7*, 7) Lum, 8) *Cldn5*. Slides were re-coverslipped and imaged as before.

Slide scans were analyzed using QuPath 0.2.3 quantitative pathology and bioimage analysis freeware. Briefly, the two scans for each slide were imported and affine transform matrices for alignment were obtained using: Analyze/Interactive image alignment/autoalign/Estimate transform; with settings: Registration type -> Affine transform, Alignment type -> Image intensity, Pixel size -> 20. Positive signal foci were identified using Analyze/Cell detection/Cell detection with the setting parameters designated in Table S2. Parameters were manually adjusted to allow for detection of multiple foci in clusters with a bias toward avoiding false positive detection events. Coordinates of detected foci in each channel were output as .tsv files using the Counting function (Counting/Convert detections to points, followed by Counting/Save points).

Because the assay lacked a counterstain that could be used to delineate cell boundaries, we relied on transcript proximity to designate positions positive for expression and coexpression. Coordinates were analyzed in R to identify transcript locations within putative expressing cells and to estimate the positions of coexpression between transcripts. Briefly, foci located within an approximate sub-cellular length (10 um) of two other foci for the same probe were considered transcripts in an expressing cell, i.e. were designated as positive points of expression. To detect coexpression a similar analysis was performed where a positive marker gene (*Gad2*, *Slc17a6*, *Pax6*, *Slc17a7*, *Lum*, or *Cldn5*) position was considered to be coexpressed with *Cyp26b1* when it lied within 10 um of two positive positions for *Cyp26b1*. For *Hoxa10*, which was more lowly expressed than the other assayed genes, a positive designation for expression was given when at least one other

Hoxa10 position fell within 10 um of a focus, and a coexpression designation required one positive *Hoxa10* position within 10um of a given positive marker gene position. Expression and coexpression mapping was compared between the RNAscope and sci-Space datasets.

3.6.19 Variance decomposition model

The variance of gene expression within and between cell populations was computed using the angular distance metric. Specifically, let y = the vector of log-scaled gene expression levels for each gene in a cell, normalized to be of unit magnitude, let $\mathbb{E}(Y)$ = the arithmetic mean of log-scaled gene expression levels normalized to be of unit magnitude, and let y = the empirical distribution of **Y** across all cells in a given analysis. Then:

$$\theta(a,b) := \frac{2}{\pi} cos^{-1}\left(\frac{A \cdot B}{||A||||B||}\right)$$

$$Var(Y) := \mathbb{E}[\theta(Y - \mathbb{E}[Y])^2]$$

This variance statistic behaves similarly to the variance of a univariate distribution. Most importantly, it obeys the Law of Total Variance **(Figure 3.46C)**. If cells sampled from **Y** are partitioned into groups **X**, then:

$$\begin{aligned} Var(Y) =& \mathbb{E}[Var(Y|X)] + Var(\mathbb{E}[Y|X]) \\ &= \text{weighted average of within-group variance} + \\ &\quad \text{variance of averages among groups} \end{aligned}$$

The "variance explained" by a grouping **X**, e.g. grouping cells by cell type, can therefore be computed as:

$$\text{Variance explained} = 1 - \frac{\text{within-group variance}}{\text{total variance}}$$

$$= 1 - \frac{\mathbb{E}[Var(Y|X)]}{Var(Y)}$$

This naive formula for variance explained is strongly affected by the number of groups in the grouping $\mathbf{X}$, similar to how the variance explained in a regression model is affected by the number of degrees of freedom in the model. If one divides a population of cells into random groups of two, variance explained will be high, even though the grouping has no biological meaning. We can correct for this by using an adjusted formula:

$$\text{Variance explained } = 1 - \frac{\mathbb{E}[Var(Y|X)]}{\mathbb{E}[Var(Y|\text{ permuted } X)]}$$

where in the denominator, the group id associated with each cell is randomly permuted.

This adjusted formula corrects for the number of groups in $\mathbf{X}$, but is confounded by another factor. The observed variance between cell gene expression vectors is a result of both biological heterogeneity and technical factors such as the sparsity of the single cell RNA-seq data. If the same biological sample is profiled in two different experiments, and the median number of UMIs per cell is N in experiment 1 and $4N$ in experiment 2, then the variance explained by the grouping $\mathbf{X}$ = cell type will be lower in experiment 1 vs. 2 due to increased sparsity. To correct for sparsity and estimate the proportion of biological variance explained by grouping $\mathbf{X}$, we use a final adjusted formula:

$$\text{Variance explained } = 1 - \frac{\mathbb{E}[Var(Y|X)] - \mathbb{E}[Var(\text{ resampled } Y|X)]}{\mathbb{E}[Var(Y|\text{ permuted } X)] - \mathbb{E}[Var(\text{resampled } Y|\text{ permuted } X)]}$$

In this formula, the "resampled Y" terms are computed by replacing each cell's gene expression

vector y with a sample from the distribution Multinomial .The distribution of the multinomial is parameterized by the two parameters, n the number of draws and p a vector of probabilities $(p_1, p_2, \ldots p_i)$ or the probability of drawing each gene. For each simulated cell from a given group, n is set to match the number of UMIs measured from a given cell and p is calculated by dividing the number of UMIs for a given gene in that cell grouping, by the total number of UMIs from that group. Resampling in this manner is a way of estimating what variance one would observe if there were no biological heterogeneity in a group of cells and the only source of observed heterogeneity was sparsity.

Figure 3.47 applies this formula to estimate the proportion of biological gene expression variance explained by cell type; the proportion of biological gene expression variance explained by cluster; the proportion of biological gene expression variance explained by spatial position, using the grouping **X** = a spatial spatial bin (non-overlapping $2 * 2$ spot squares); the proportion of biological gene expression variance explained by spatial position and cell type where **X** = the combination of spatial bin and cell type. **Figure 3.4B** applies this formula to estimate the proportion of the residual biological gene expression variance, after accounting for cell type, that is explained by spatial position. In these models, the grouping $\mathbf{X} = (\text{cell type, spot id})$, and permuted $\mathbf{X} = (\text{cell type, permuted spot id})$. By permuting the spot id but not the cell type annotation, we ensure that our estimate is of the variance explained by the cell type + space model relative to the cell-type-only model, rather than relative to a null model. To estimate the variance that can be explained by a null (shuffled) model, we perform this permutation procedure 50 times and take the average of these trials.

3.6.20 Pairwise angular distance

Only cell types with 100 cells or more, originating from a single slide were considered for this analysis. Each cell was size factor normalized, and scaled to the unit hypersphere. The pairwise angular distance was then calculated as detailed above. To test whether there was a relationship between angular distance and physical distance, a linear model was fit with angular distance as

the response and physical distance as the sole predictor (angular.distance distance). Reported p-values indicate the significance of coefficient for the physical distance predictor variable using the Wald linear regression test.

3.6.21 Spatial gene modules

Gene module analysis in **Figure 3.48** comparing cell type derived gene modules and spatial gene modules was performed on Slide 14. Briefly, we performed PCA, UMAP dimensionality reduction and clustering on nuclei mapping to Slide 14. Genes autocorrelated in the UMAP embedding were then calculated using Monocle3's graph_test() function. Similarly, genes autocorrelated in spatial-position were calculated using the sci-Space derived spatial coordinates as the input into Monocle3's graph_test() function. We performed module analysis on the union of genes with significant autocorrelation in the UMAP embedding and spatial position (FDR < 0.05). UMAP gene modules were recovered using Monocle3's find_gene_modules() function on the cell*gene matrix. Spatial gene modules were recovered using Monocle3's find_gene_modules() on the position*gene matrix. Gene expression from each set of discovered modules was then aggregated using Monocle 3's aggregate_gene_expression() function and visualized as row- and column-clustered heatmaps.

3.6.22 Neuronal trajectory analysis

Neuronal trajectories mapping to the brain and classified as either Radial Glia, Neurons or Glial Cells were subset from the neural lineage partition. PCA, followed by UMAP dimensionality reduction was performed on this subset. Monocle3 was then used to learn a principle graph on the UMAP embedding. For pseudotime inference, the root of the trajectory was selected by manually identifying principal graph nodes occupied by radial glia. Each trajectory was then scaled for display purposes. These nuclei were then mapped to brain regions based on segmentation performed using the Allen Institute's Anatomical Reference Brain Atlas (www.atlas.brain-map.org). To calculate pseudotime dependent genes, nuclei mapping to the Pallium, Sub Pallium and Midbrain were first subsetted. For each brain region and the trajectory contained within it, pseudotime

dependent genes were then recovered by fitting a natural spline with three degrees of freedom using pseudotime as the predictor. Genes with at least one significant knot (FDR < 0.01) were deemed pseudotime dependent. This process was repeated for all three trajectories. The results of these tests are provided in File S3. The intersection of significant genes across all three trajectories is shown in **Figure 3.5F** . For gene ontology analysis, the dendrogram producing the row-clustering in **Figure 3.5F** was cut to produce 4 groups. Each group was then provided as input into http://geneontology.org. Displayed gene ontology terms were chosen from a list of significant terms (FDR < 0.01).

3.7 Supplemental Figures

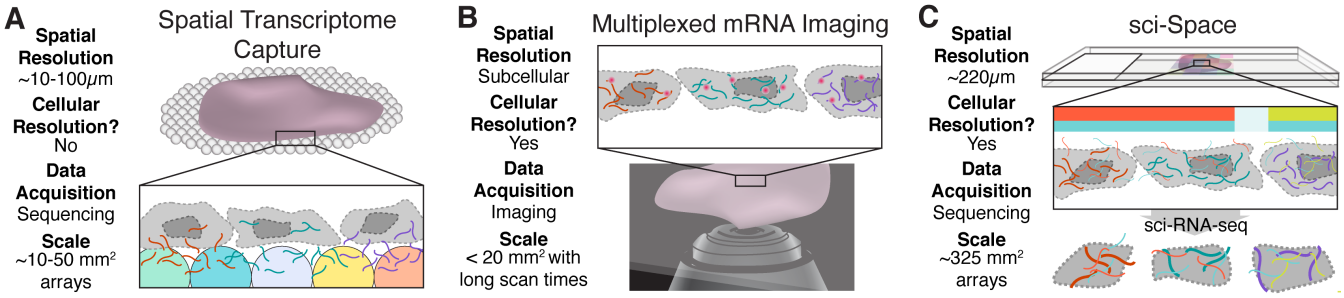

Figure 3.6: **Comparison of methods for spatial transcriptomics.** Schematics and key attributes of the two major classes of contemporary spatial transcriptomics methods, as well as sci-Space. **(A)** Spatial transcriptome capture (STC) methods, e.g. the original "spatial transcriptomics" method (*93*) and Slide-seq (*94*). **(B)** Methods relying on multiplexed mRNA imaging, e.g. MERFISH (*95*), seqFISH (*96*), and FISSEQ (*97*). **(C)** sci-Space.

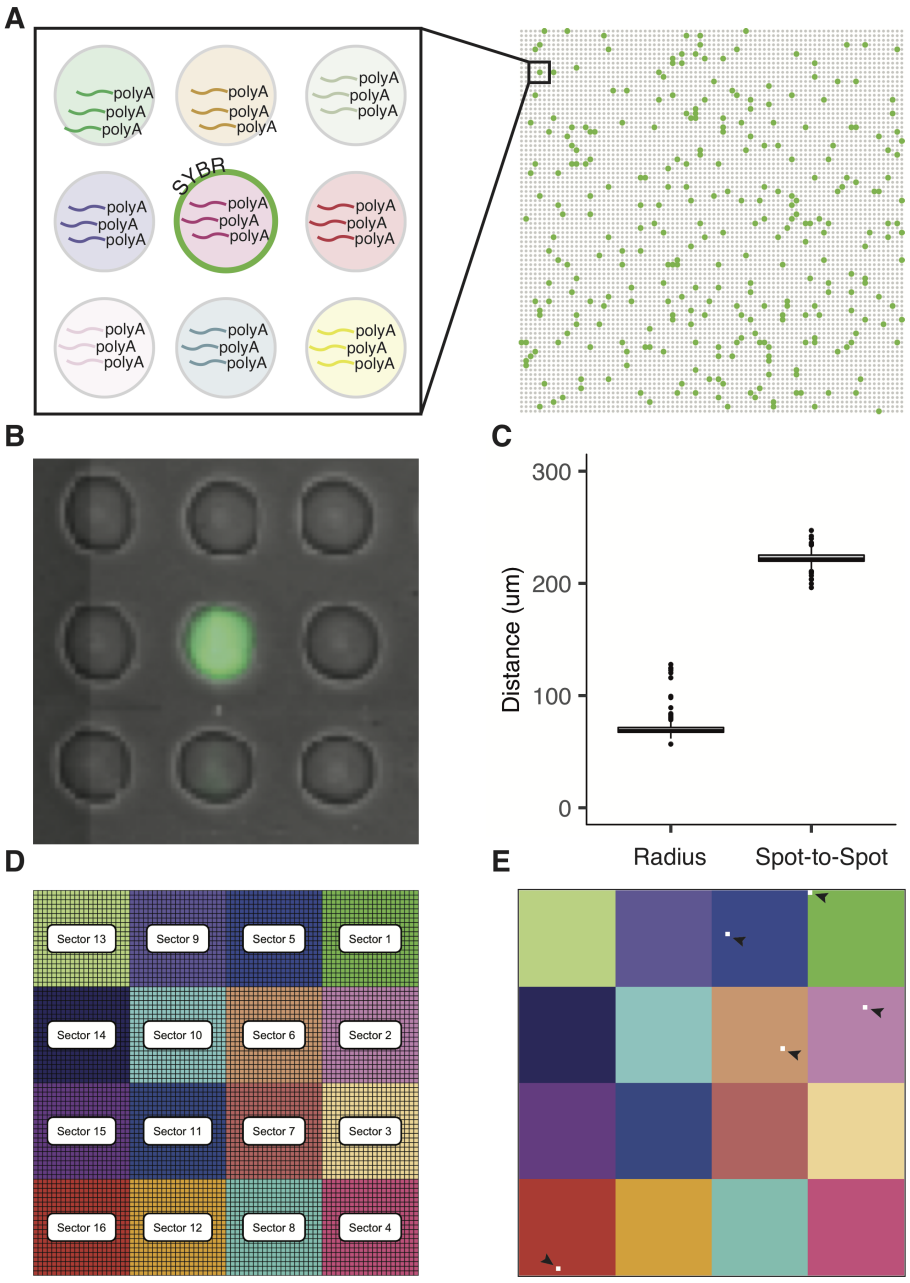

Figure 3.7: **Comparison of methods for spatial transcriptomics. (A)** Schematic of spotted oligos with SYBR green fluorescent dye marked positions labeled in green. All positions contain a combination of location-informative "hashing" sequences (single stranded DNA) with polyA tails. **(B)** Overlay of bright-field and fluorescence image of the same position. **(C)** Average radius and spot-to-spot distance computed from imaged slides. (D) Diagram of hierarchical barcoding approach where each position is marked by a unique combination of one of 16 sector barcodes (colors) and one of 1536 spot sequences. **(E)** An example displaying a single spot oligo barcode (white square) which is in 5 different sectors. Scale bar **(B)** = 0.1 mm.

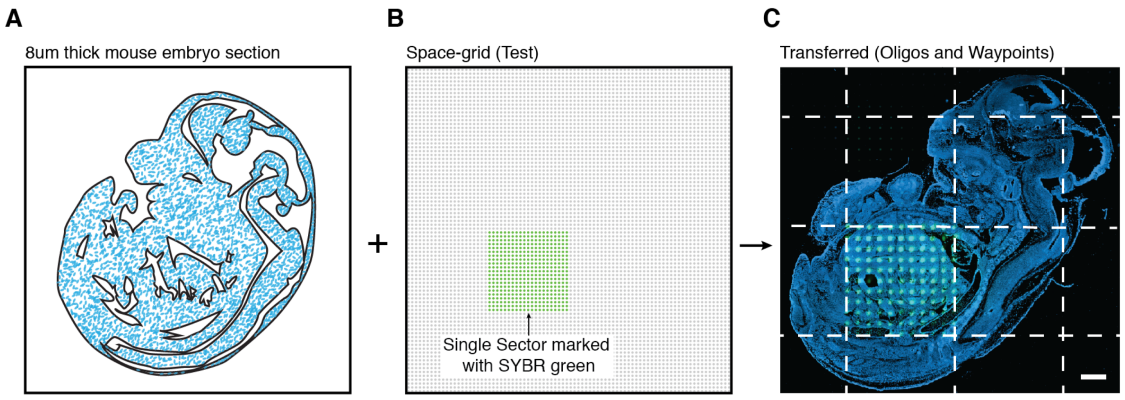

Figure 3.8: **SYBR green waypoints transfer to DAPI stained embryo. (A)** Permeabilized mouse embryo section receives **(B)** SYBR green waypoints spotted at a single section. **(C)** The resulting transfer and imaging shows the location of each waypoint on the DAPI stained section. Dashed white lines denote the approximate location of each sector. Scale bar in panel **(C)** = 0.5 mm.

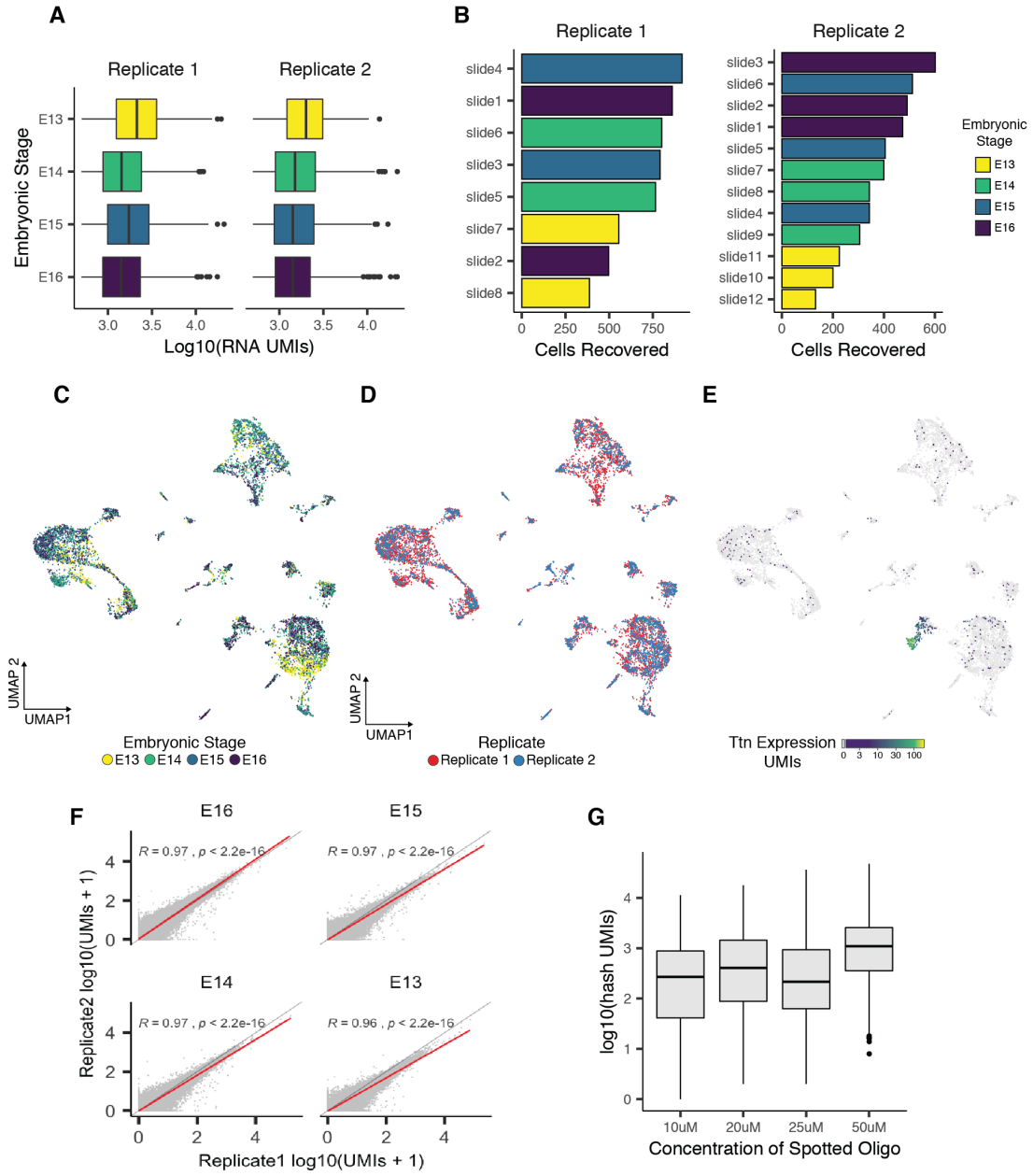

Figure 3.9: **Labeling of cryosectioned tissues with hash oligos from an agarose coated slide is compatible with sci-RNA-seq.** Slides from sections of the developing mouse embryo were first labeled with a slide specific oligo and then labeled with another hash oligo from a space-grid containing a single hash oligo at varying concentrations. Replicates are independent experiments performed on different days using tissue sections from a single batch. **(A)** RNA UMIs recovered per cell across stages and replicates. **(B)** Number of cells sampled from each slide across stages and replicates. **(C-E)** UMAP embedding colored by **(C)** embryonic stage, **(D)** replicate or **(E)** expression of skeletal muscle marker Titin (Ttn). **(F)** Correlation of RNA UMIs recovered per gene between replicates at different stages. **(G)** Hash UMIs recovered per cell of oligo spotted at 10μM, 20μM, 25μM and 50μM concentrations.

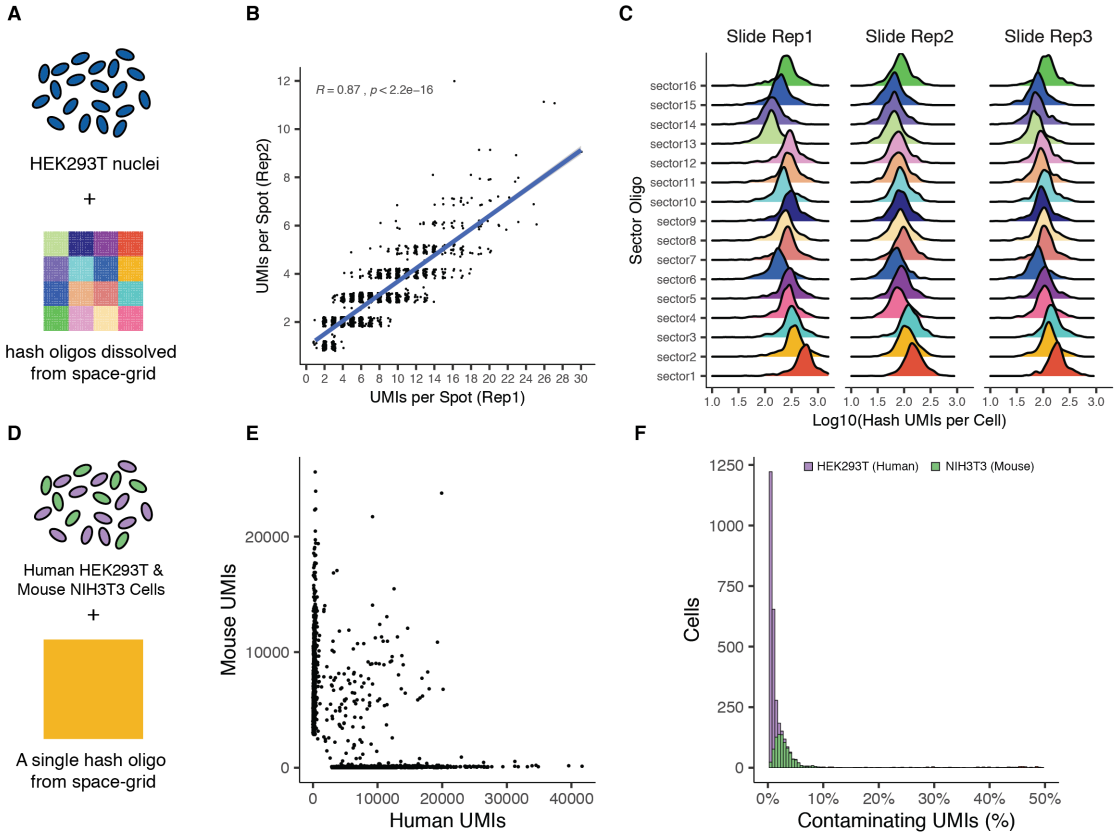

Figure 3.10: **Spotted space-grids are reproducible.** **(A)** HEK293T nuclei were exposed to hash-oligos dissolved from one of 3 space-grids. **(B)** Correlation between spot oligo counts originating from different slides. **(C)** Distribution of sector oligos observed per cell, broken out by replicate. **(D)** Cartoon depicting control experiment with human and mouse cell lines grown on a glass slide and barcoded with a single hash-oligo. **(E)** Scatter plot depicting the number of human (X axis) or mouse (Y axis) unique molecular identifiers (UMIs) detected per cell. Nuclei were filtered for those mapping to a single slide. **(F)** The percentage of contaminating UMIs for each cell type.

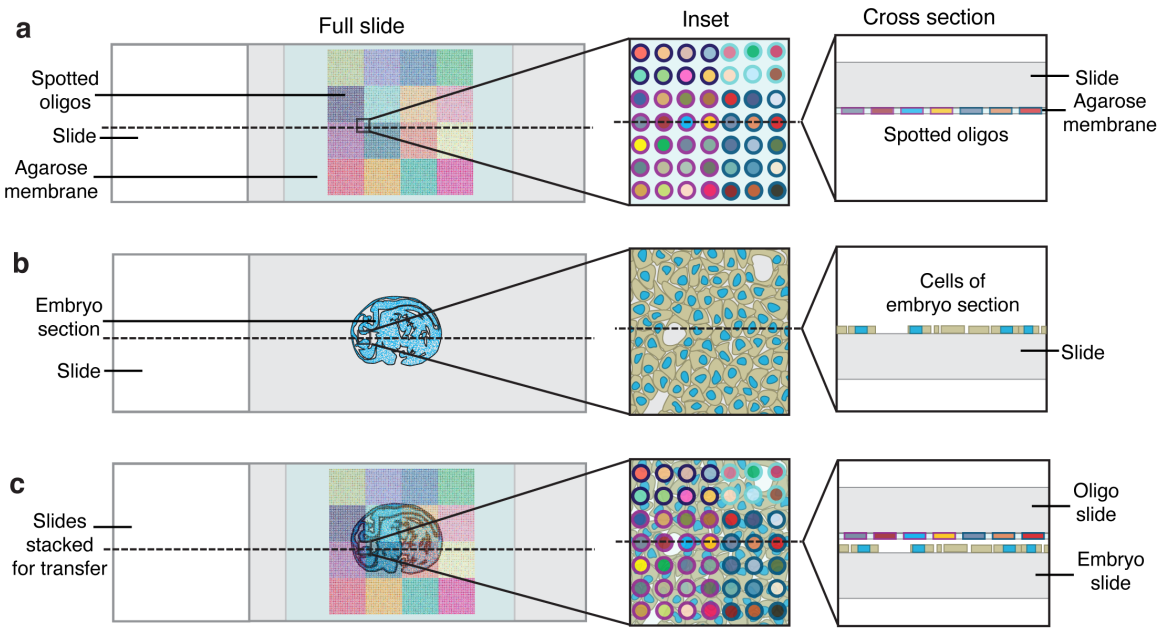

Figure 3.11: **Spotted space-grids are reproducible.** **(A)** Spatially indexed slides, "space-grids," were fabricated by spotting unique combinations of hashing oligos onto agarose membrane-coated slides. **(B)** Permeabilized fresh-frozen tissue sections **(C)** received the spatially-defined pattern of oligos by diffusion from the space-grids when the oligo-laden agarose and tissue section were sandwiched together between their carrier slides.

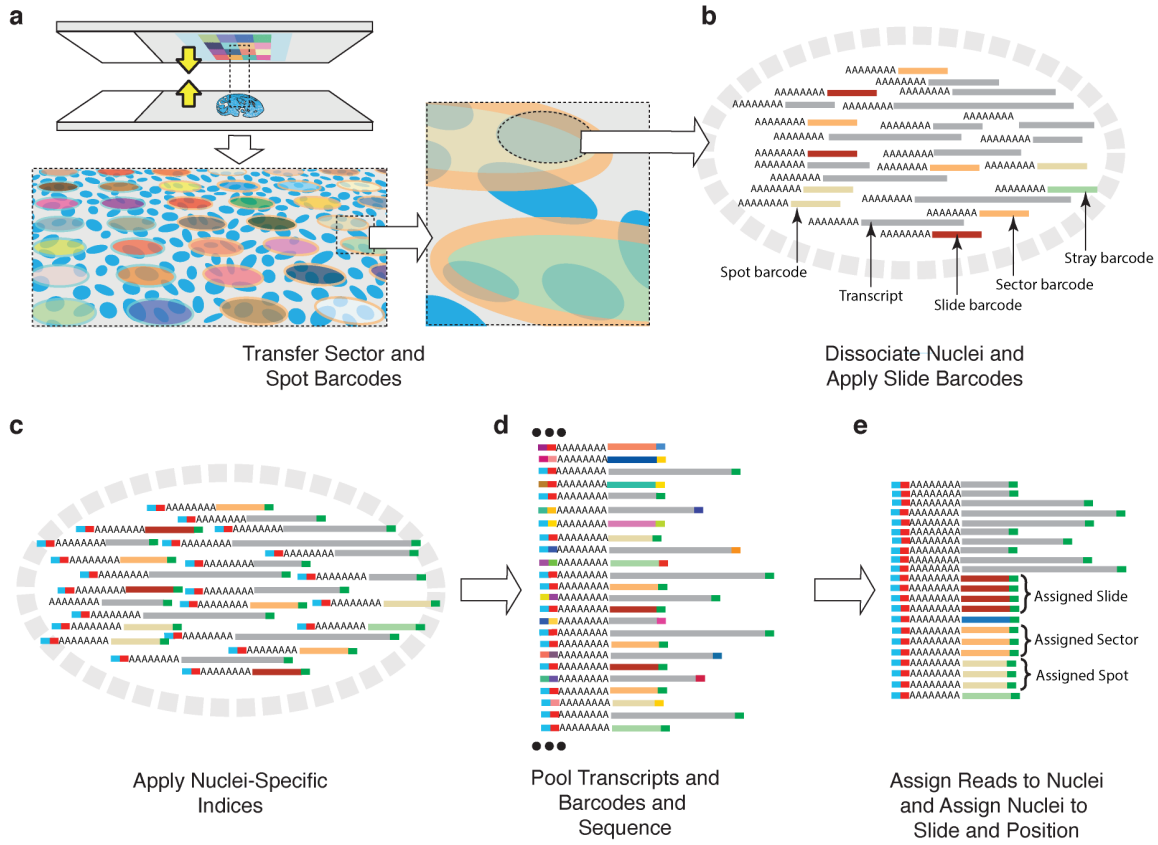

Figure 3.12: **sci-Space workflow for sequencing library preparation and demultiplexing transcripts allows transcripts and spatial positions to be assigned to individual nuclei.** **(A)** Hashing oligos or barcodes are transferred to nuclei as determined by nuclei positions relative to the barcode array. **(B)** Nuclei from each slide are dissociated and labelled with an additional slide-specific barcode. **(C)** Transcripts and barcodes are tagged with nuclei-specific indices according to the sci-RNA-seq protocol (*120*). Note that the green segment shown in the schematic includes combinatorial indexing barcodes introduced during sci-RNA-seq. **(D)** Barcodes and transcripts from all nuclei are pooled and sequenced. **(E)** Indices are used to demultiplex transcripts and barcodes, which allow for the assignment of each nucleus to its slide, sector, and spot of origin.

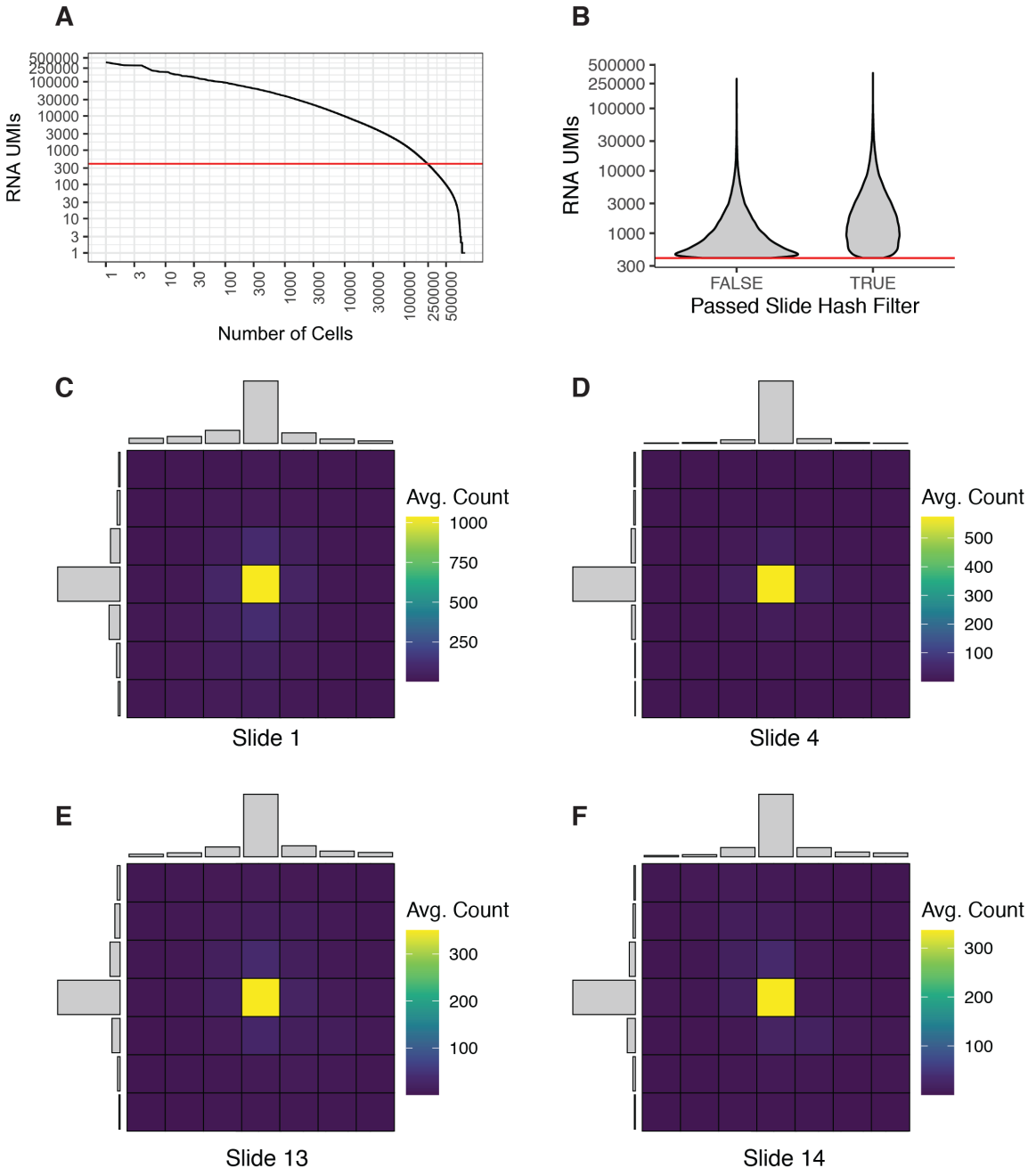

Figure 3.13: **Cellular hashing distinguishes low RNA UMI nuclei from aggregates and uniquely marks a nucleus' position.** **(A)** Cumulative distribution of RNA UMIs (unique molecular identifiers) identified after sequencing. Red line corresponds to a lenient cutoff used for initial calls. **(B)** Violin density plots displaying the distribution of cells which were labeled by single slide-specific hash oligo. Cells which failed to show enrichment of a single slide-specific hash oligo were filtered out. Red line corresponds to the same cutoff displayed in panel **(A)**. **(C-F)** Heatmaps displaying the average hash oligo counts from the top spot and surrounding positions for 1000 randomly sampled cells from Slide 1 **(C)**, Slide 4 **(D)**, Slide 13 **(E)** and Slide 14 **(F)**. Histograms reflect the marginal sums for rows or columns.

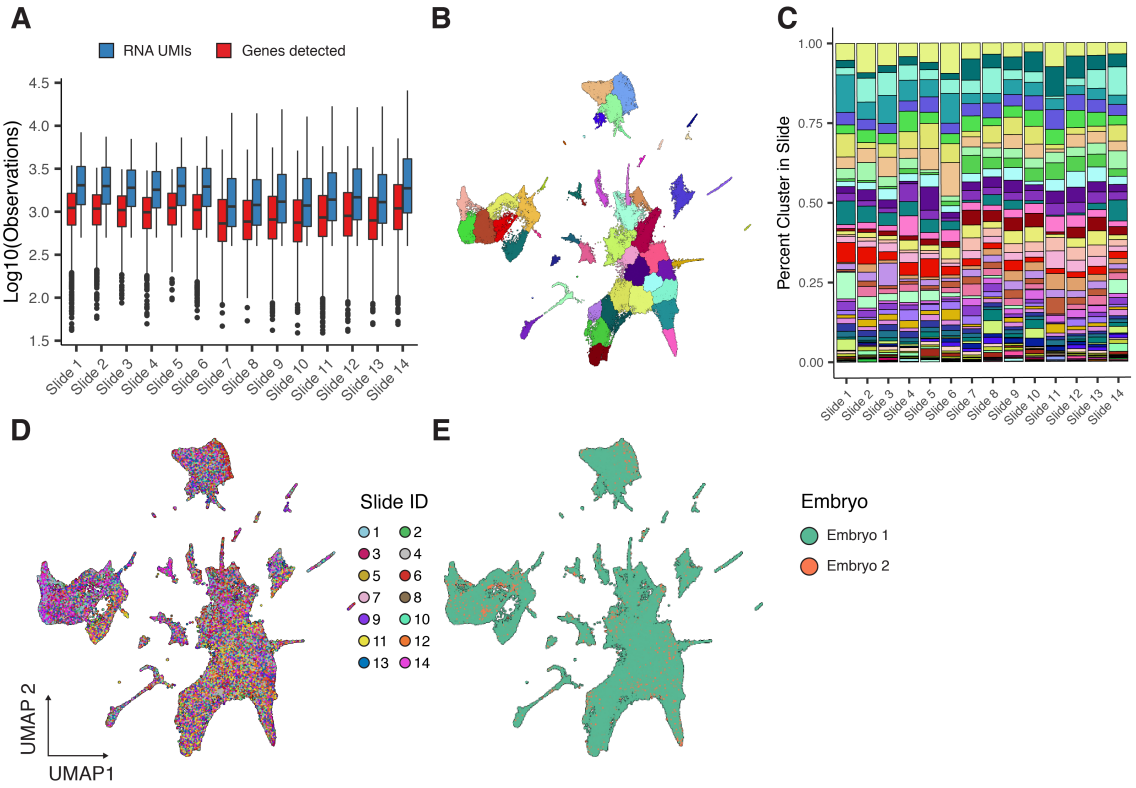

Figure 3.14: **sci-Space sequenced cells have complex transcriptomes and separate into the major cell types.** **(A)** Boxplots displaying the number of unique molecular identifiers recovered (red) and genes detected (blue) for cells from each slide. **(B-C)** Louvain clustering result and the corresponding proportions of cells found in each cluster. **(D-E)** UMAP embedding colored by the slide of origin **(D)** or embryo of origin **(E)**.

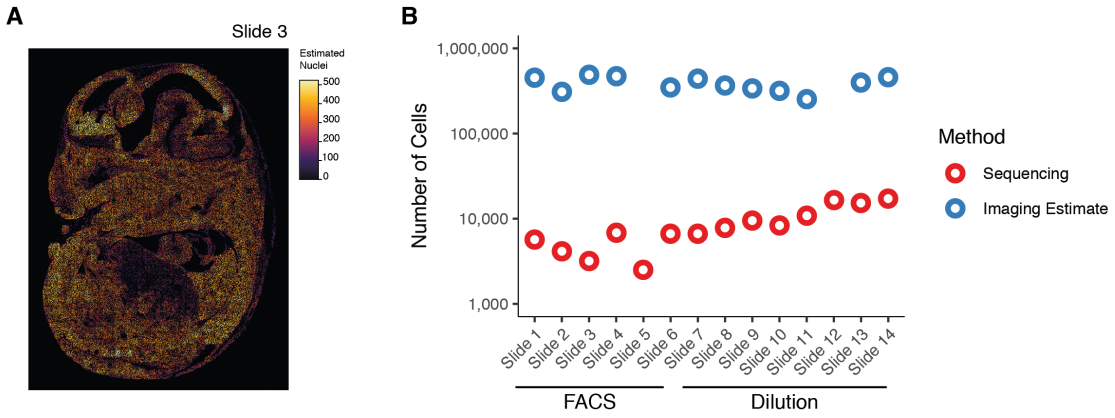

Figure 3.15: **Comparison of recovery of nuclei from sequencing and estimated nuclei present upon imaging.** **(A)** Estimates of DAPI stained nuclei present per spot. Per spot estimates for Slide 3 are shown as an example. **(B)** Comparison of the total number of estimated nuclei present versus the number of nuclei recovered from each slide. Method for seeding cells during sci-RNA-seq is noted below the x-axis. An estimate of nuclei count from slides 5 and 12 could not be computed due to corruption or loss of the image file.

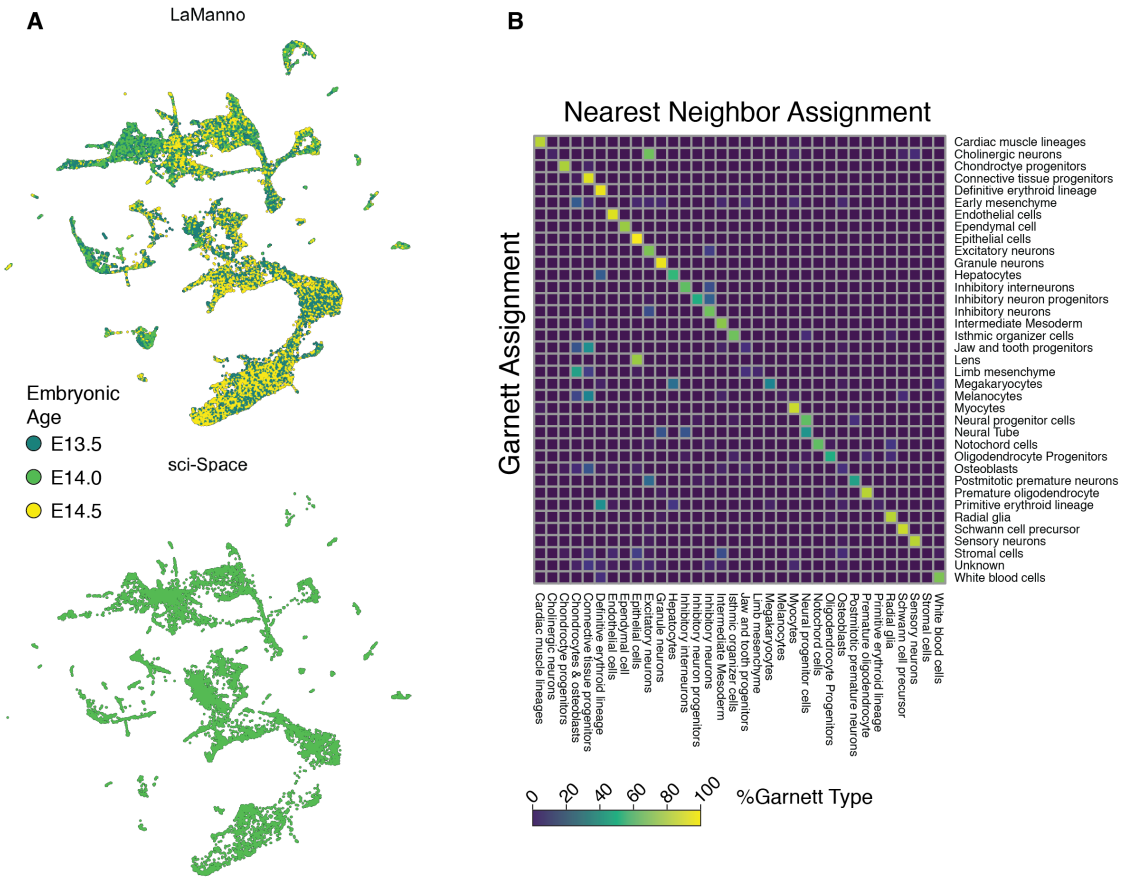

Figure 3.16: **Automated cell type annotation and concordance between methods.** **(A)** Estimates of DAPI stained nuclei present per spot. Per spot estimates for Slide 3 are shown as an example. **(B)** Comparison of the total number of estimated nuclei present versus the number of nuclei recovered from each slide. Method for seeding cells during sci-RNA-seq is noted below the x-axis. An estimate of nuclei count from slides 5 and 12 could not be computed due to corruption or loss of the image file.

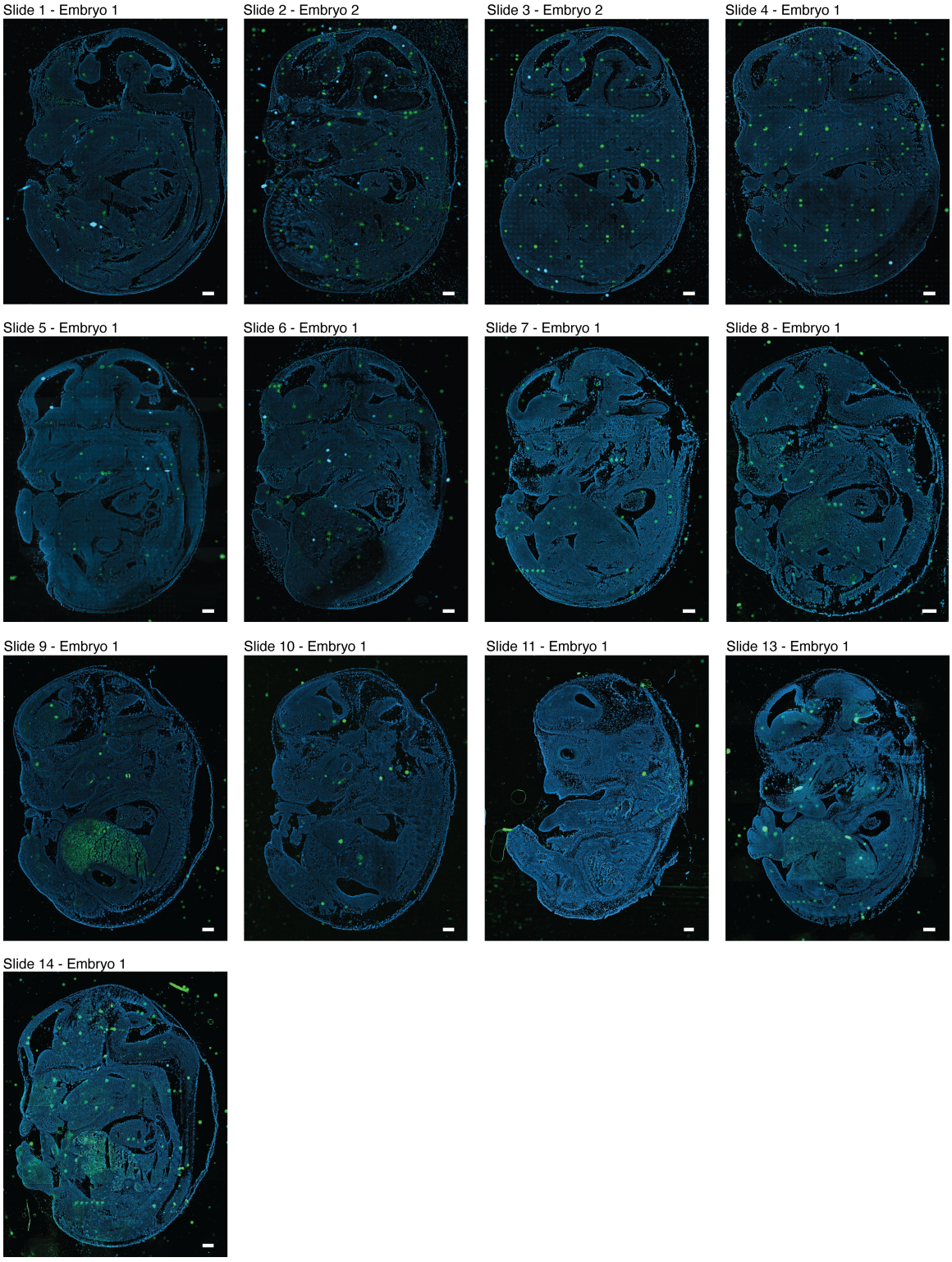

Figure 3.17: **DAPI stained images of sci-space sequenced slides.**The slide number and the embryo from which they originate are displayed above each image. The image for Slide 12 was lost during data transfer. SYBR green point layouts varied between space-grid prints, and positioning of the sections relative to the oligo array also varied. Scale bars = 0.5 mm.

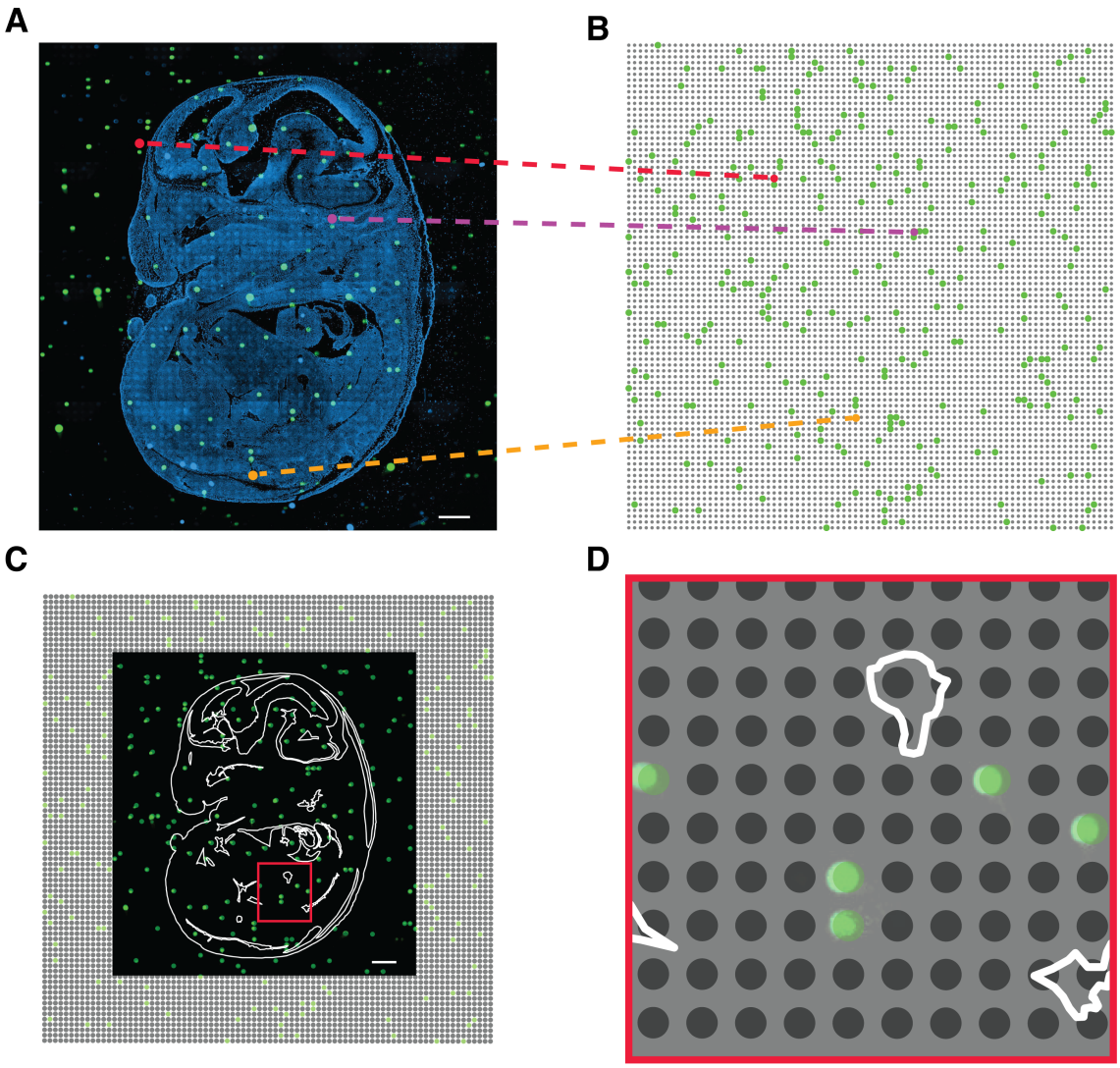

Figure 3.18: **Co-registration procedure of imaged section and space-grid.** **(A)** DAPI stained E14.0 section (Slide 3) with SYBR green points imaged in the GFP-channel. Matched SYBR green waypoints between the image and **(B)** the intended SYBR pattern on an ideal space-grid are used to calculate an affine-transformation. **(C)** Co-registered imaging data with inferred positions overlayed with image with inset highlighted **(D)**. Scale bars in panels **(A)** and **(C)** = 1 mm.

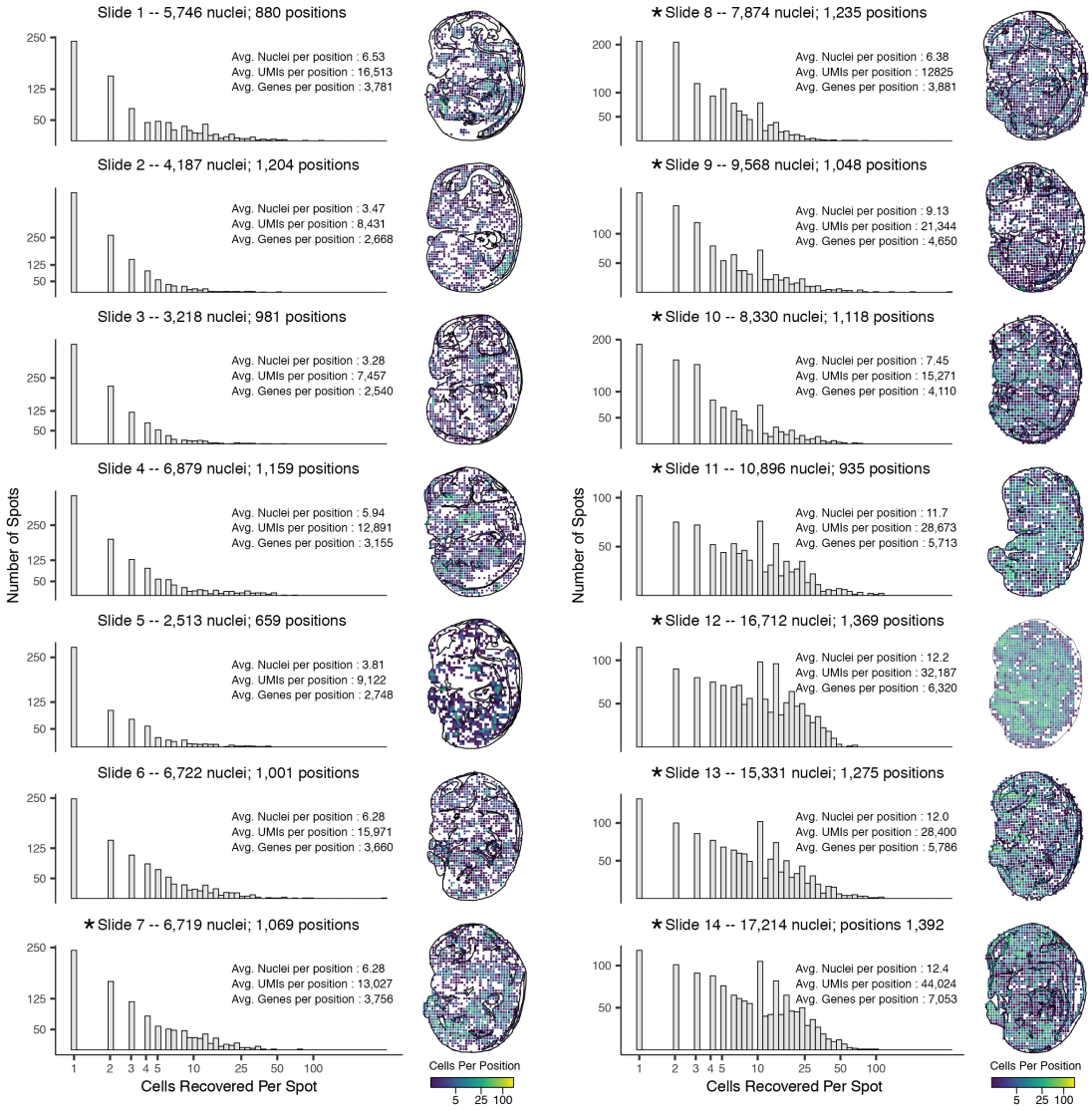

Figure 3.19: **Spatial distribution of nuclei per slide in addition to key spatial metrics.**Histograms displaying the number of nuclei recovered per position for each slide are shown next to a heatmap of the cells recovered per spatial position. In text, above and beside each histogram, summary spatial statistics are noted for each slide. An asterisk is displayed next to slides that were processed using the optimized sci-Space protocol with an extra sonication step and nucleus seeding via dilution (as opposed to FACS).

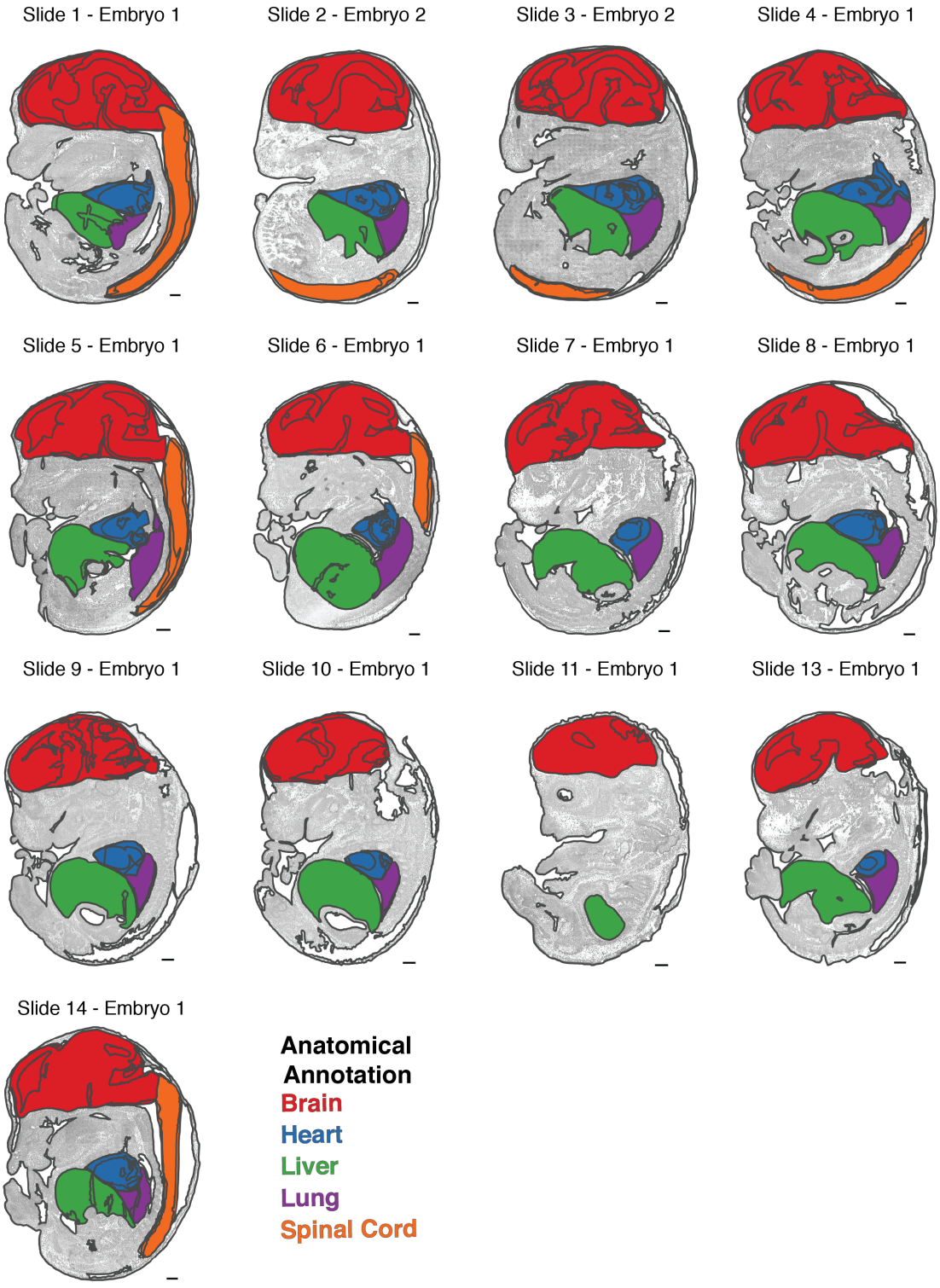

Figure 3.20: **Segmented organs from each embryo section.** Segmented regions are highlighted on each section and colored according to the denoted anatomical structure. 500um scale bar shown in the bottom right corner of each image.

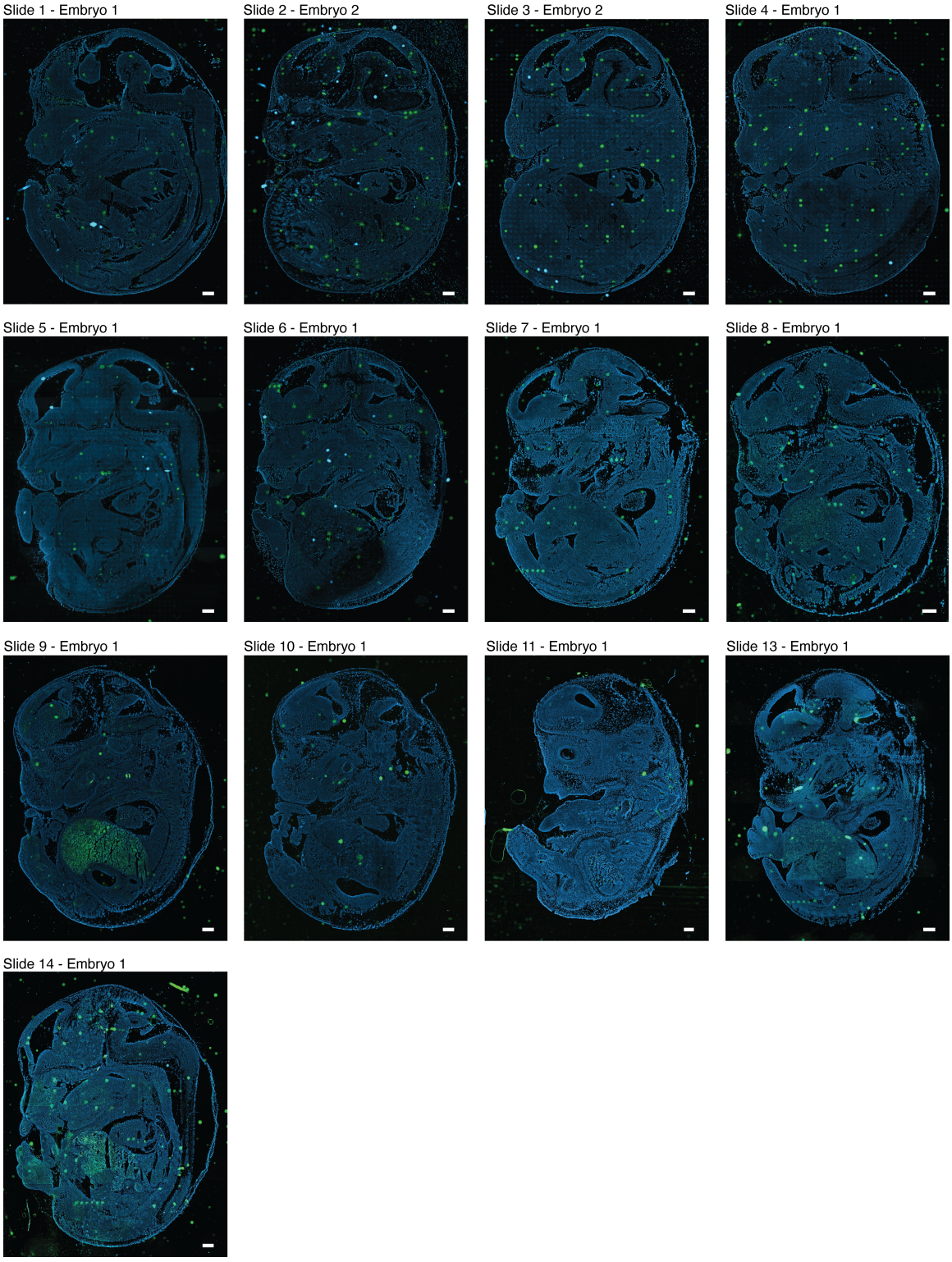

Figure 3.21: **Alignment of immunostaining from adjacent cryosections.** **(A-D)** Serial cryosections adjacent to the sequenced embryo tissue sections immunostained for sarcomeric alpha-actinin (green), cardiac alpha-myosin heavy chain (red), and E-cadherin (magenta) with a Hoechst (blue) counterstain. **(E-H)** Affine transform alignment allowed for overlay of the staining on the image of the sequenced and for identification of tissue structures, as seen in **(I,J)**, the cardiac region insets from panels **(D,H)** respectively. Scale bars: (A-H) = 1 mm, (I, J) = 0.5 mm.

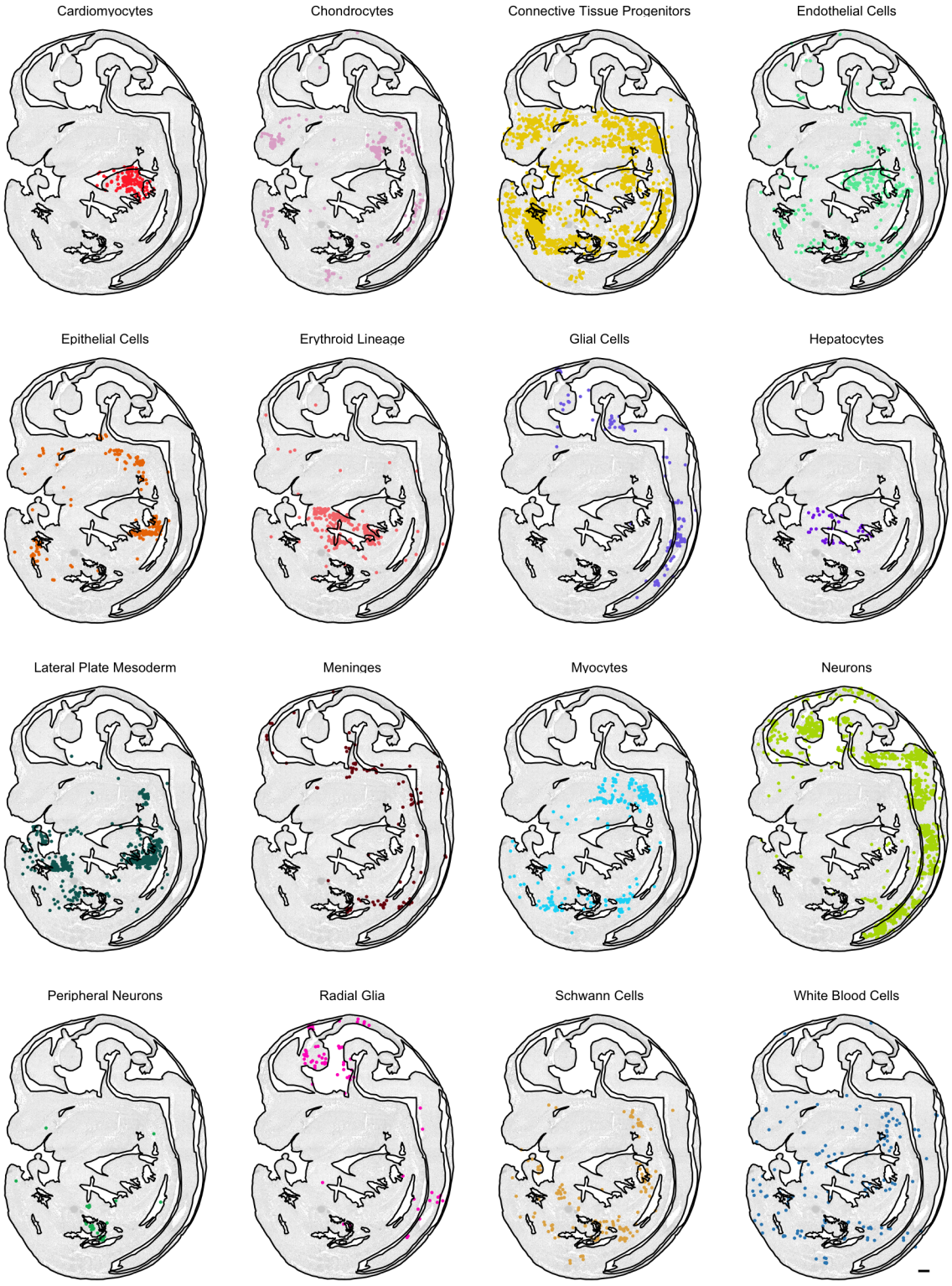

Figure 3.22: **Spatial position of cell types for Slide 1 from embryo .** Sequenced cells, broken out by cell type, are placed onto the imaged embryo at the spatial position to which they map on top. Top 16 abundant cell types are shown with a 500um scale bar (bottom right).

Figure 3.23: **Spatial position of cell types for Slide 4 from embryo 1.** Sequenced cells, broken out by cell type, are placed onto the imaged embryo at the spatial position to which they map on top. Top 16 abundant cell types are shown with a 500um scale bar (bottom right).

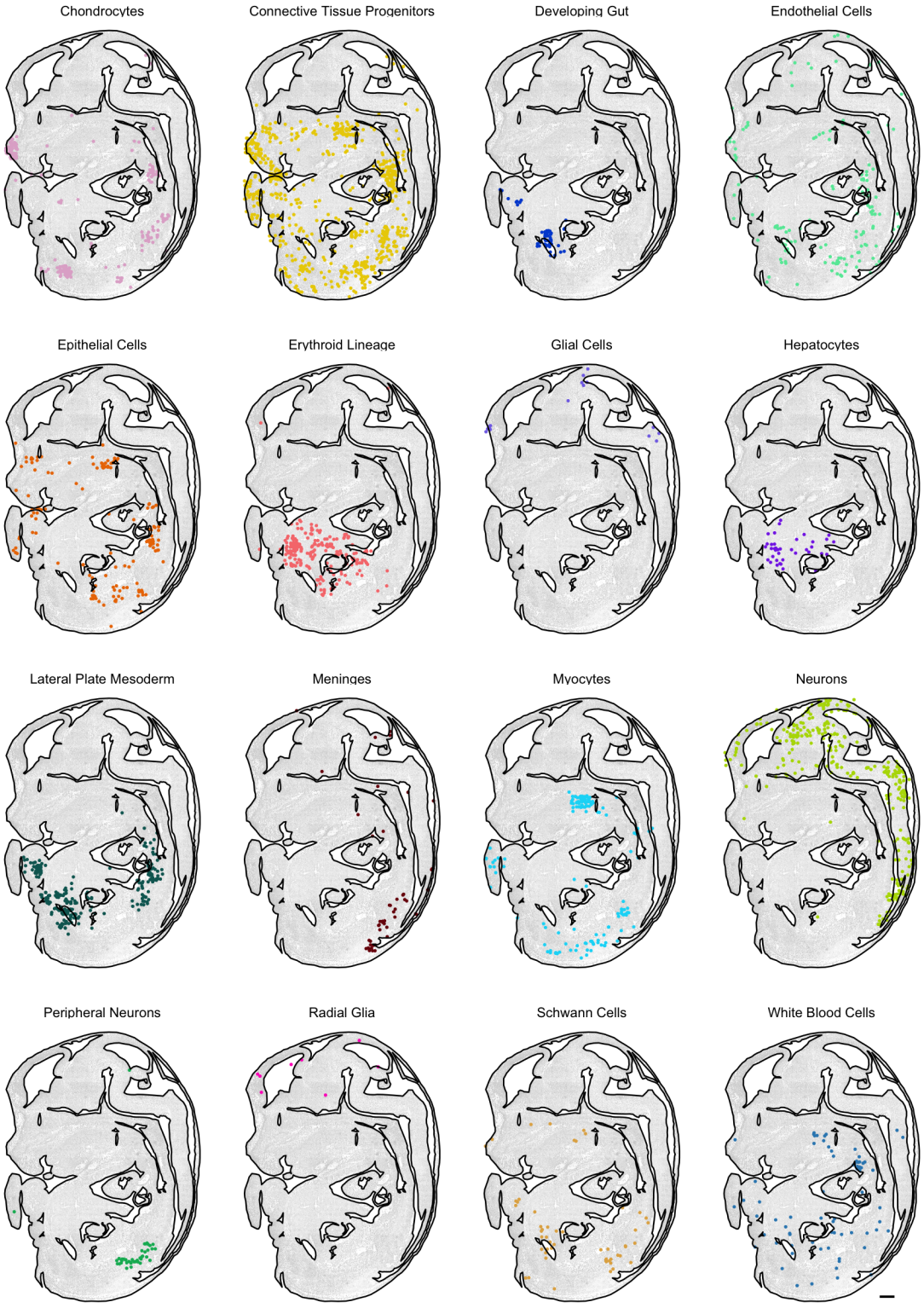

Figure 3.24: **Spatial position of cell types for Slide 5 from embryo 1.** Sequenced cells, broken out by cell type, are placed onto the imaged embryo at the spatial position to which they map on top. Top 16 abundant cell types are shown with a 500um scale bar (bottom right).

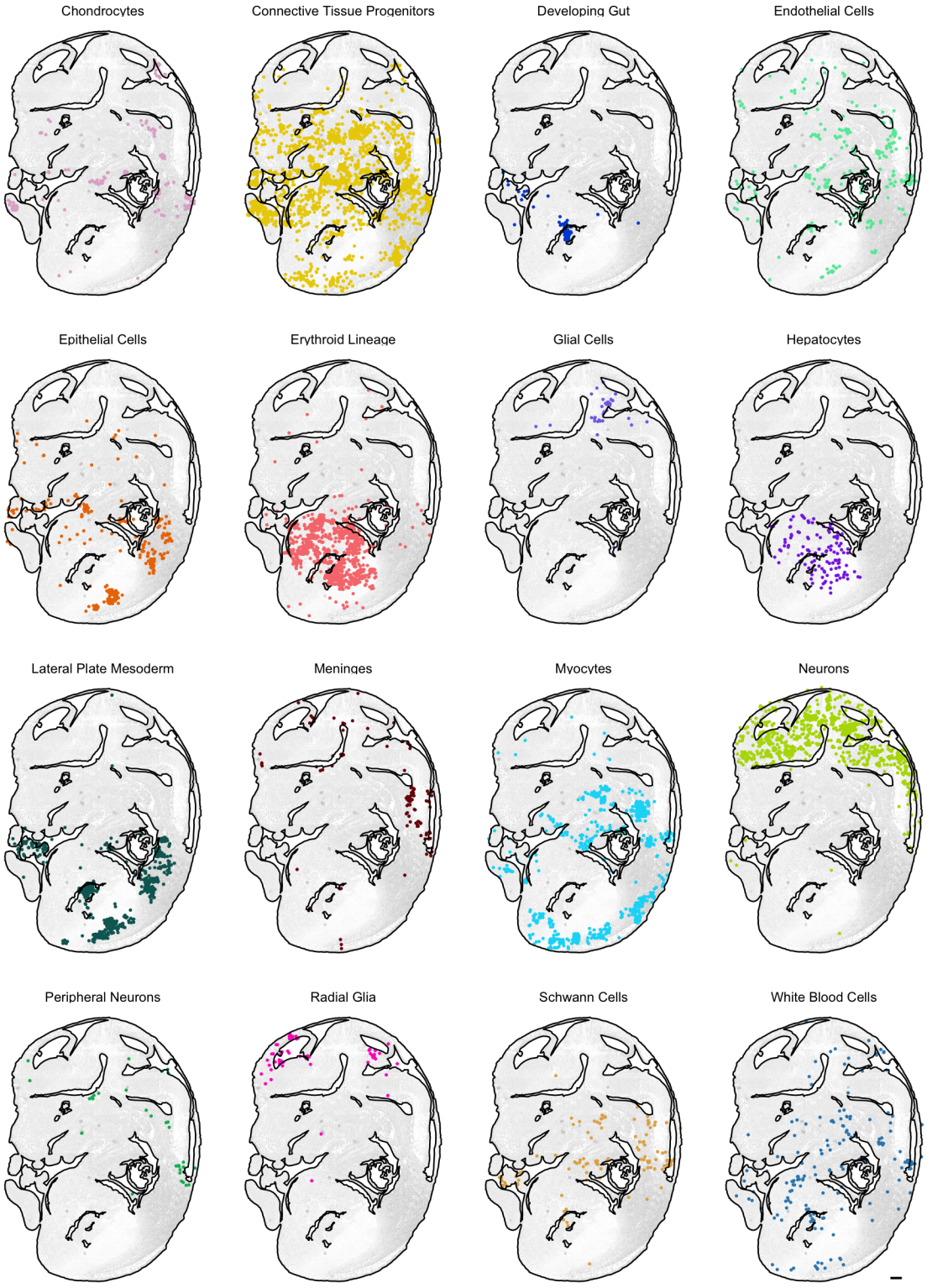

Figure 3.25: **Spatial position of cell types for Slide 6 from embryo 1.** Sequenced cells, broken out by cell type, are placed onto the imaged embryo at the spatial position to which they map on top. Top 16 abundant cell types are shown with a 500um scale bar (bottom right).

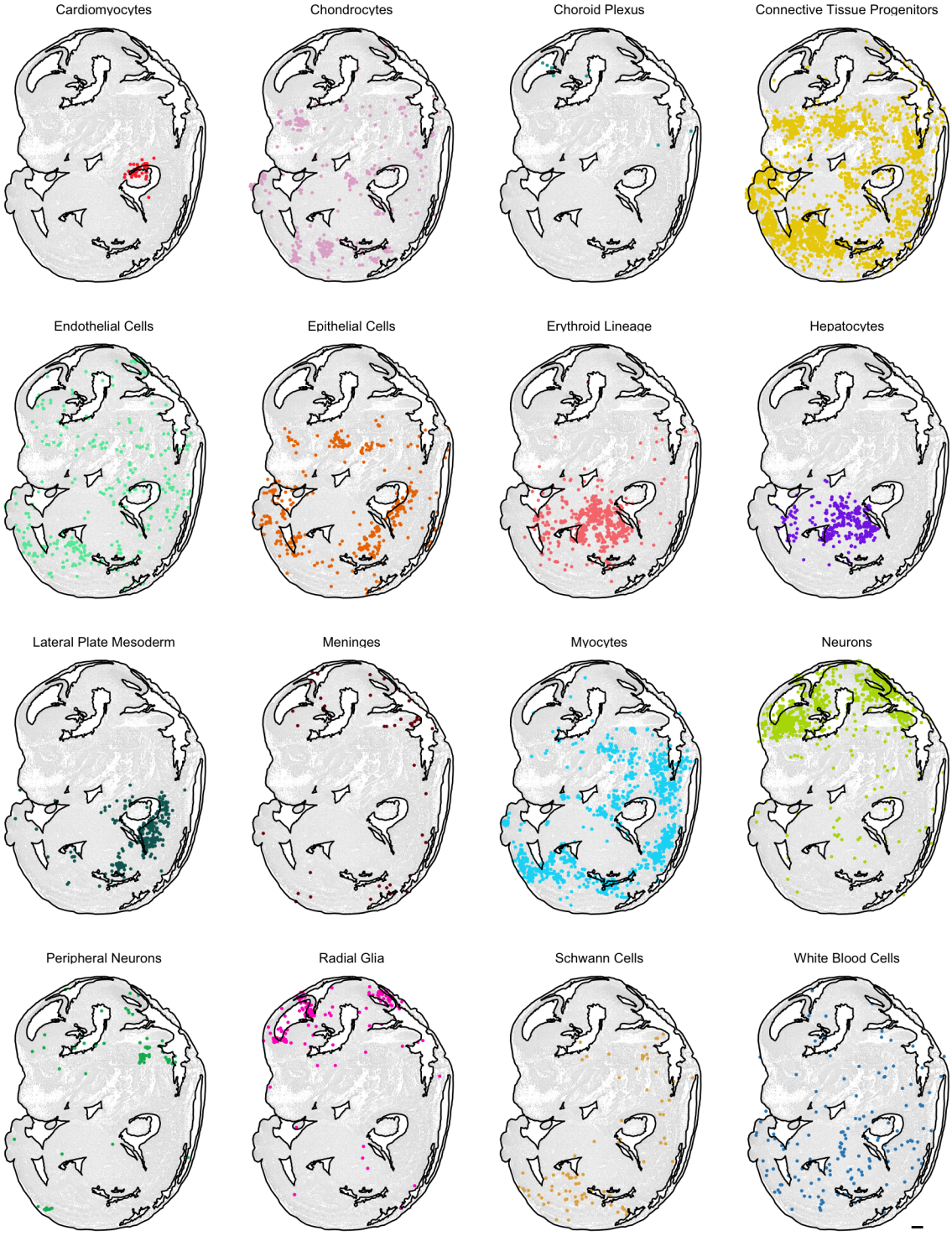

Figure 3.26: **Spatial position of cell types for Slide 7 from embryo 1.** Sequenced cells, broken out by cell type, are placed onto the imaged embryo at the spatial position to which they map on top. Top 16 abundant cell types are shown with a 500um scale bar (bottom right).

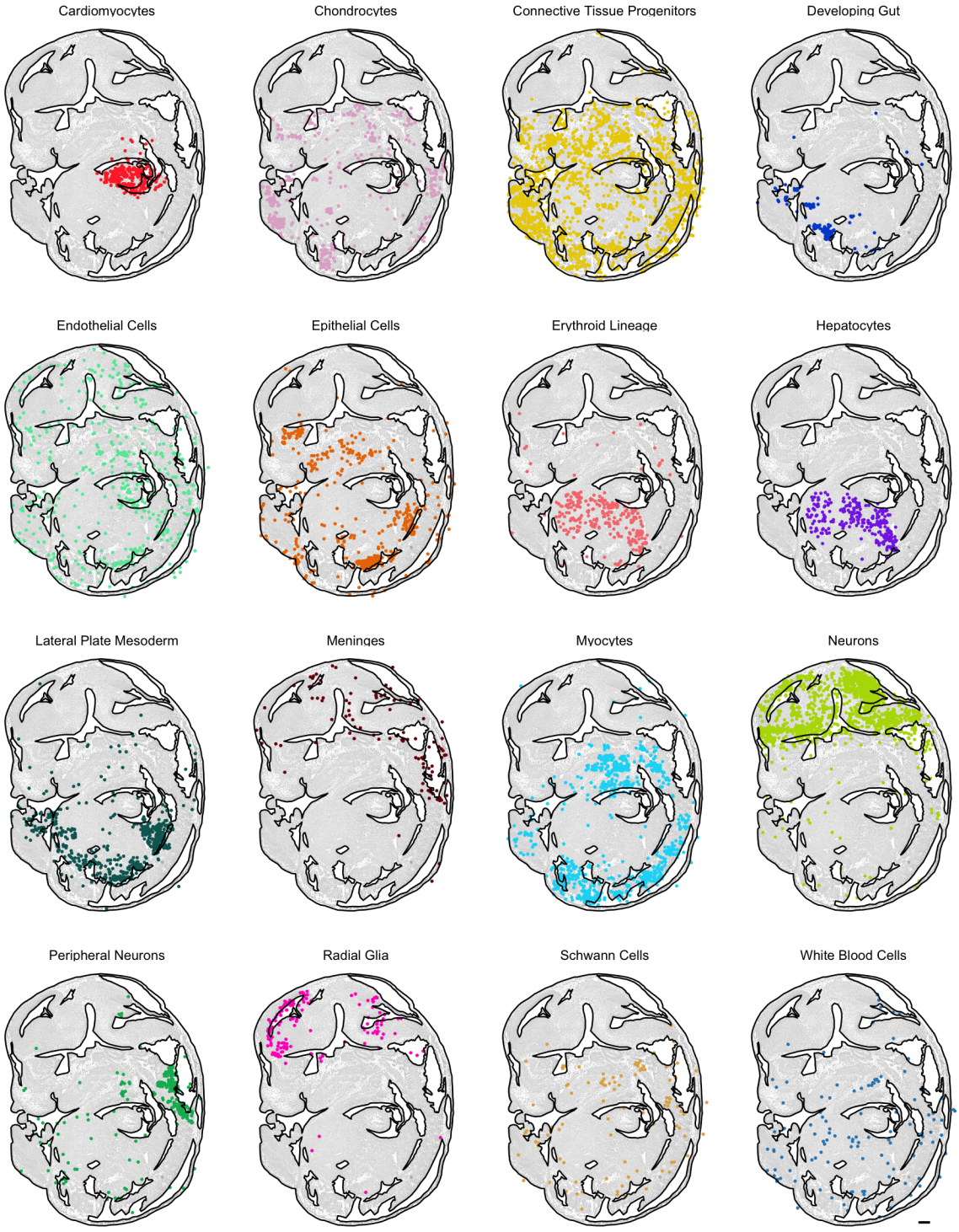

Figure 3.27: **Spatial position of cell types for Slide 8 from embryo 1.** Sequenced cells, broken out by cell type, are placed onto the imaged embryo at the spatial position to which they map on top. Top 16 abundant cell types are shown with a 500um scale bar (bottom right).

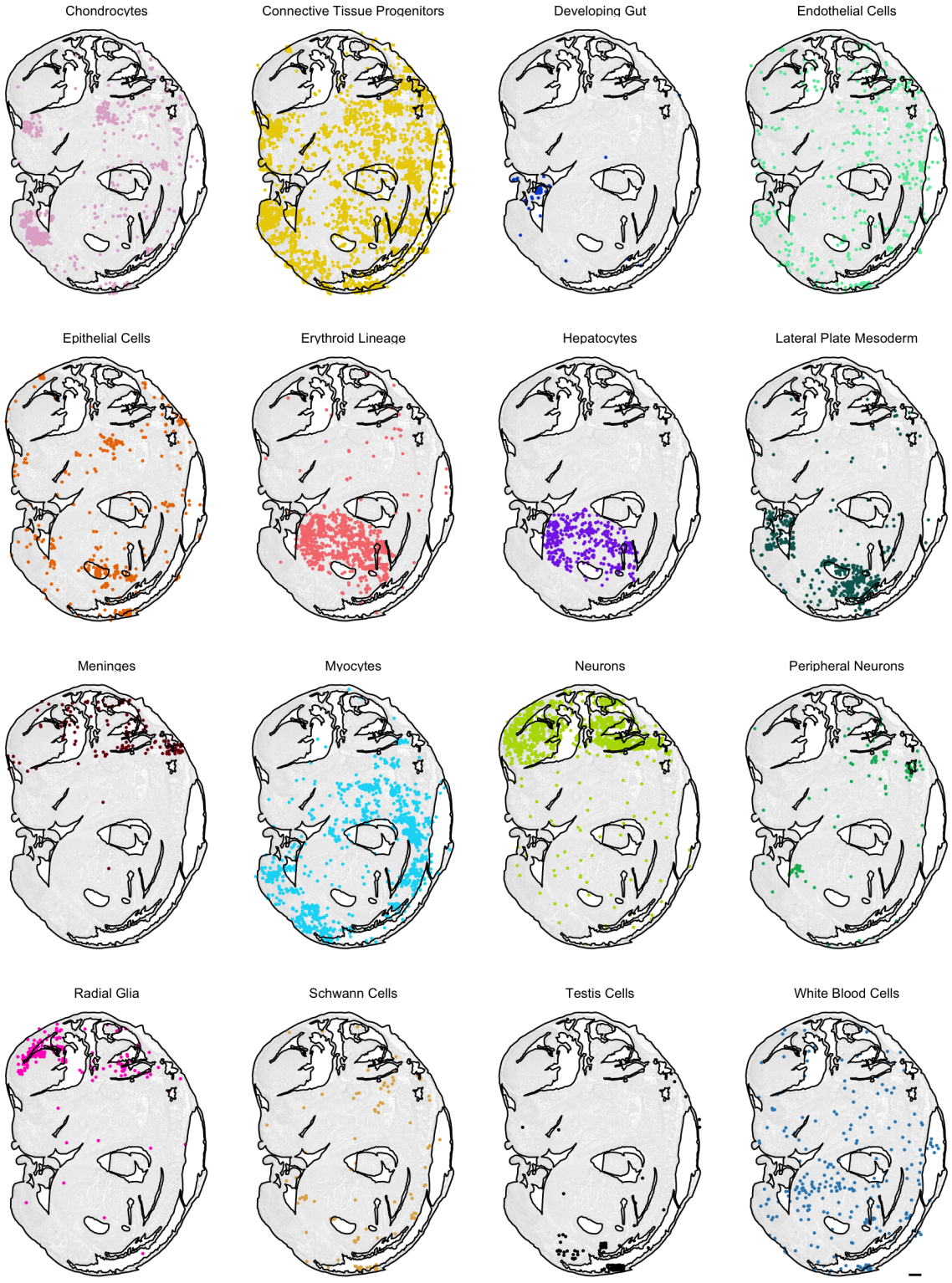

Figure 3.28: **Spatial position of cell types for Slide 8 from embryo 1.** Sequenced cells, broken out by cell type, are placed onto the imaged embryo at the spatial position to which they map on top. Top 16 abundant cell types are shown with a 500um scale bar (bottom right).

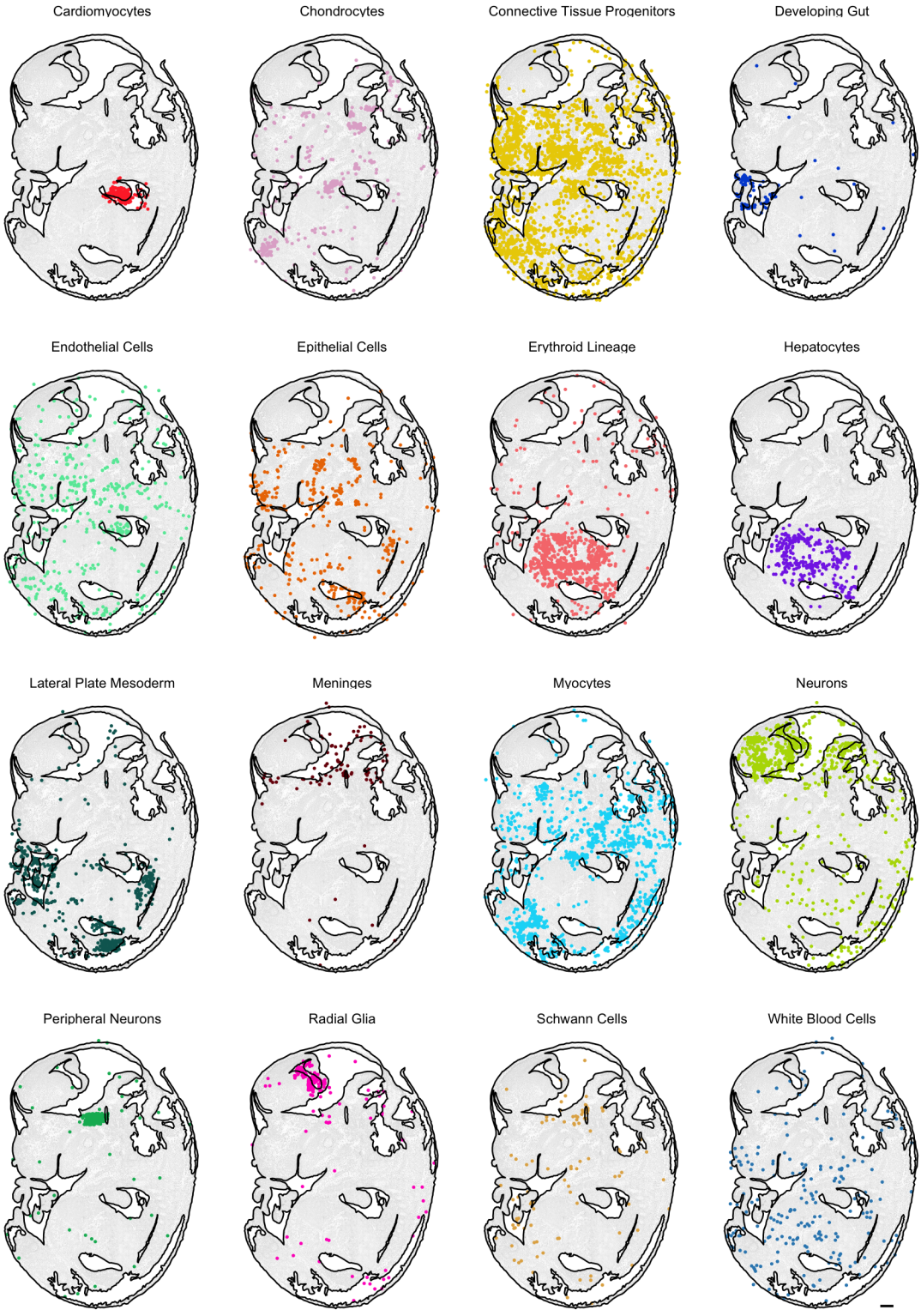

Figure 3.29: **Spatial position of cell types for Slide 8 from embryo 1.** Sequenced cells, broken out by cell type, are placed onto the imaged embryo at the spatial position to which they map on top. Top 16 abundant cell types are shown with a 500um scale bar (bottom right).

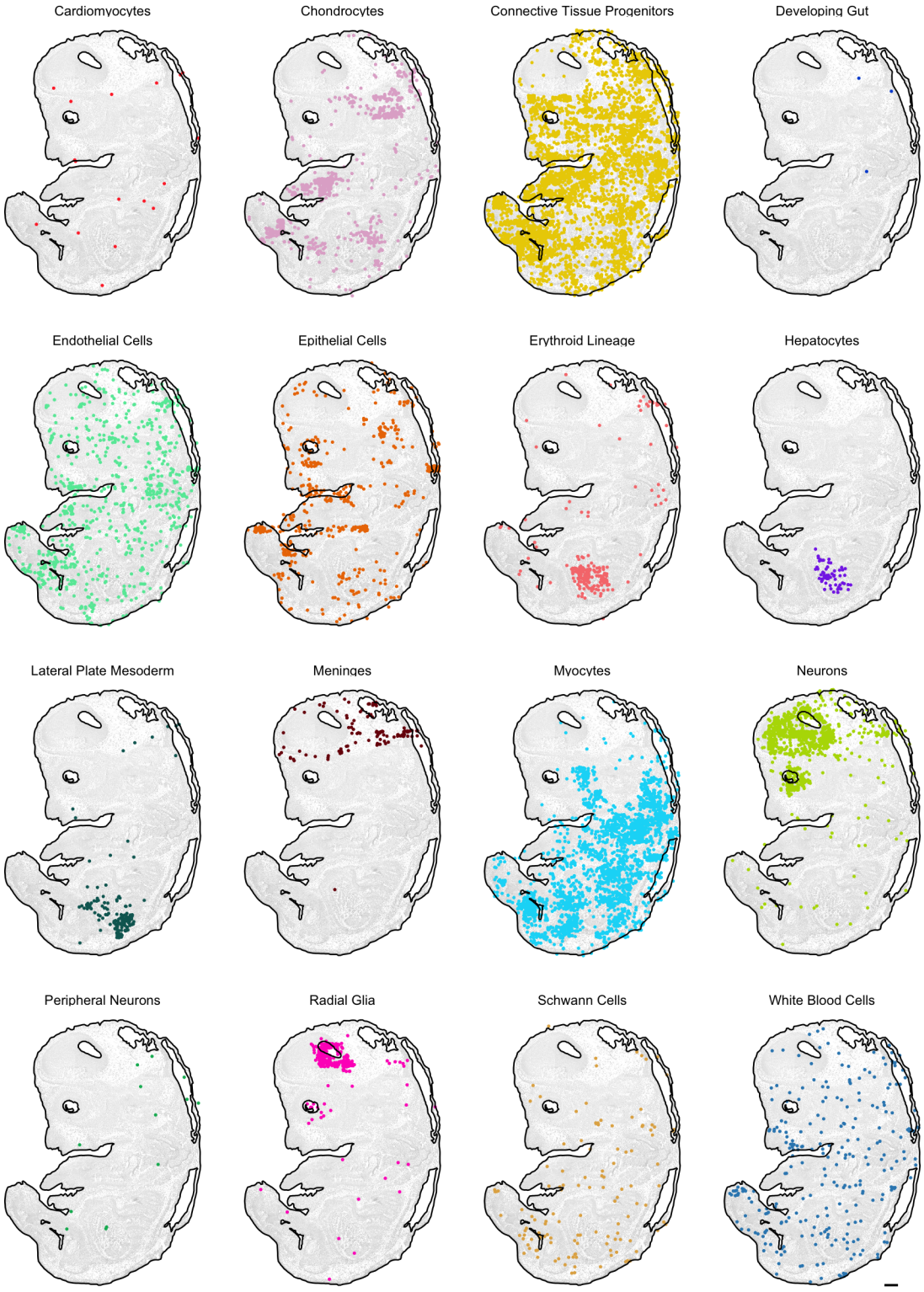

Figure 3.30: **Spatial position of cell types for Slide 8 from embryo 1.** Sequenced cells, broken out by cell type, are placed onto the imaged embryo at the spatial position to which they map on top. Top 16 abundant cell types are shown with a 500um scale bar (bottom right).

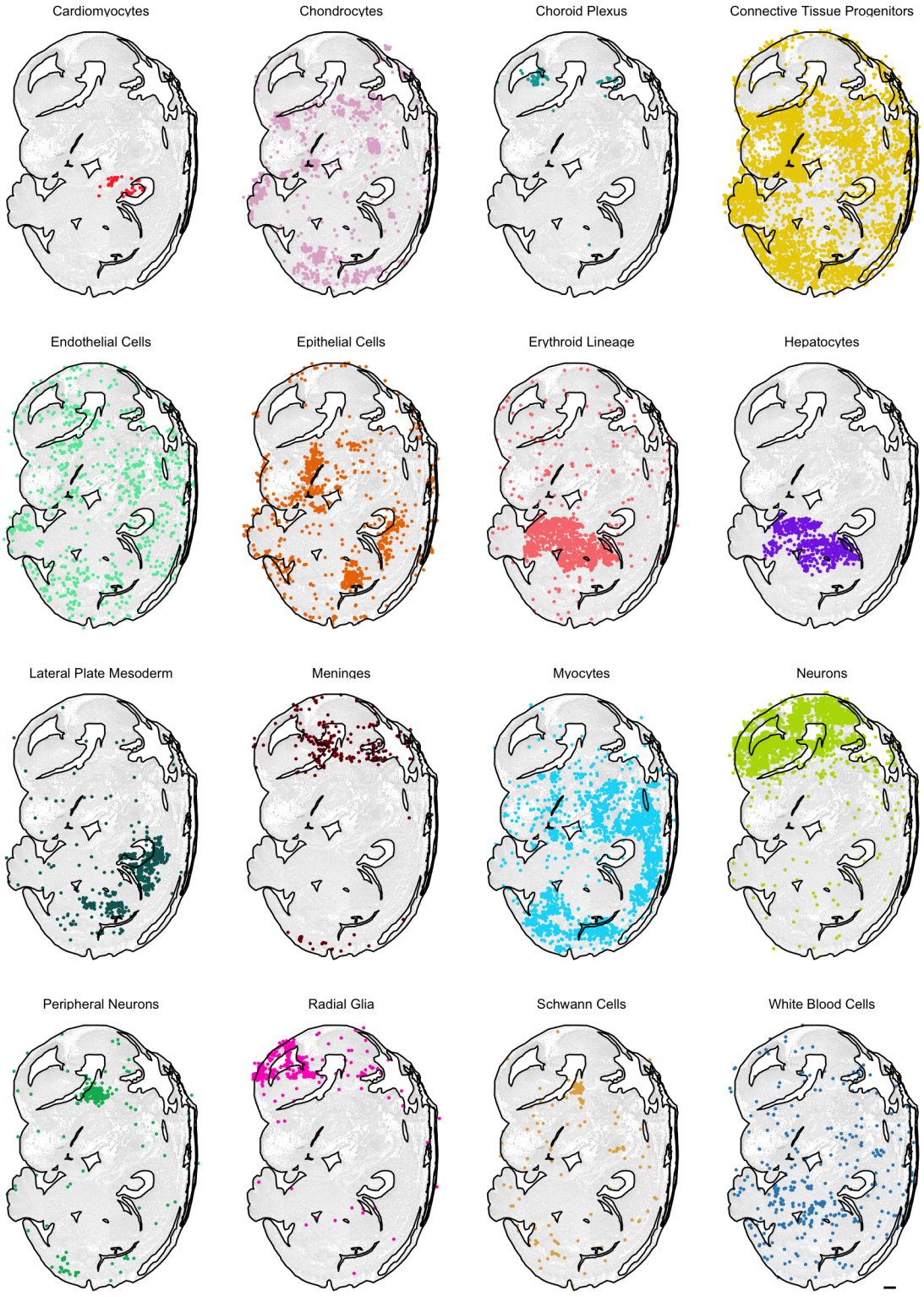

Figure 3.31: **Spatial position of cell types for Slide 8 from embryo 1.** Sequenced cells, broken out by cell type, are placed onto the imaged embryo at the spatial position to which they map on top. Top 16 abundant cell types are shown with a 500um scale bar (bottom right).

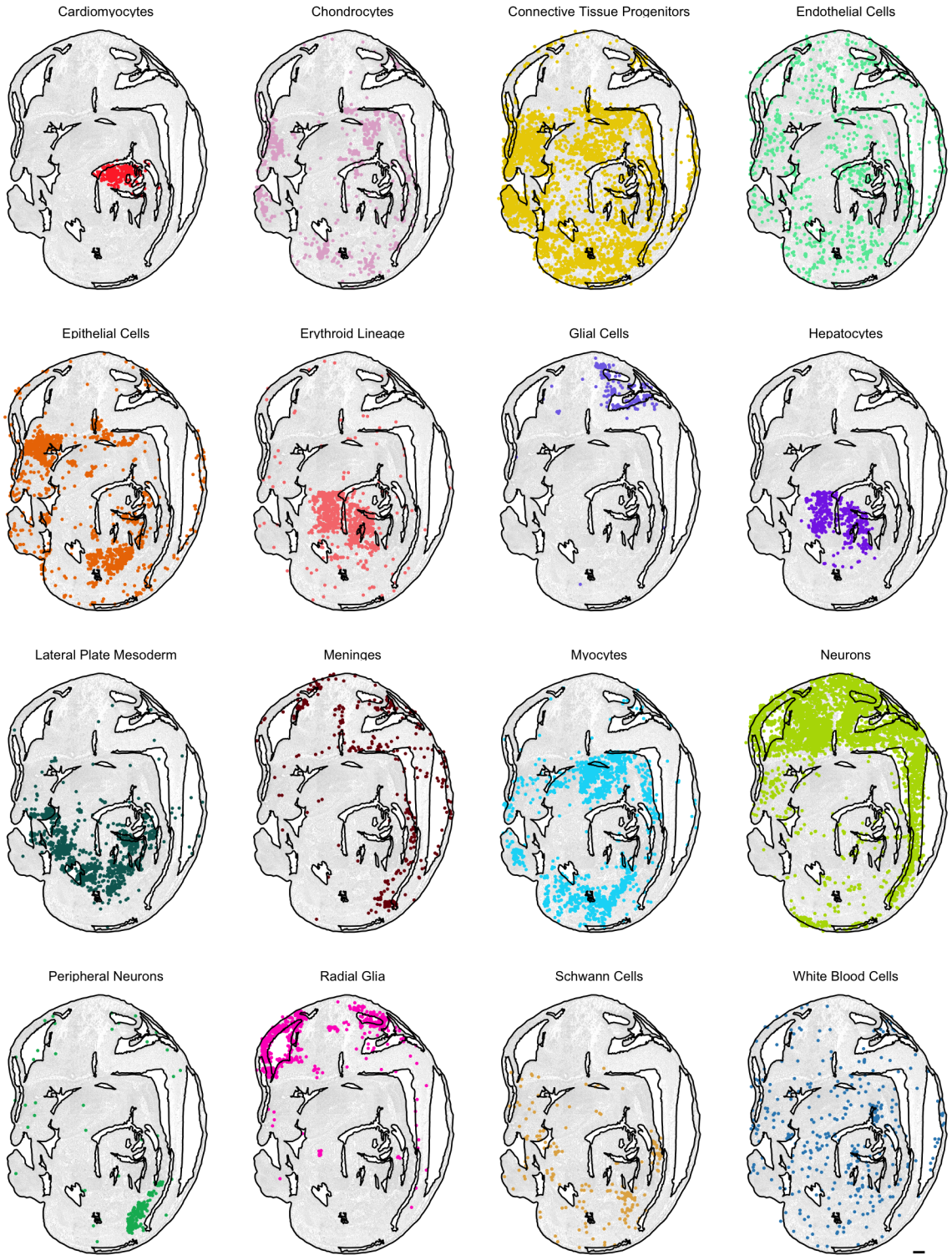

Figure 3.32: **Spatial position of cell types for Slide 8 from embryo 1.** Sequenced cells, broken out by cell type, are placed onto the imaged embryo at the spatial position to which they map on top. Top 16 abundant cell types are shown with a 500um scale bar (bottom right).

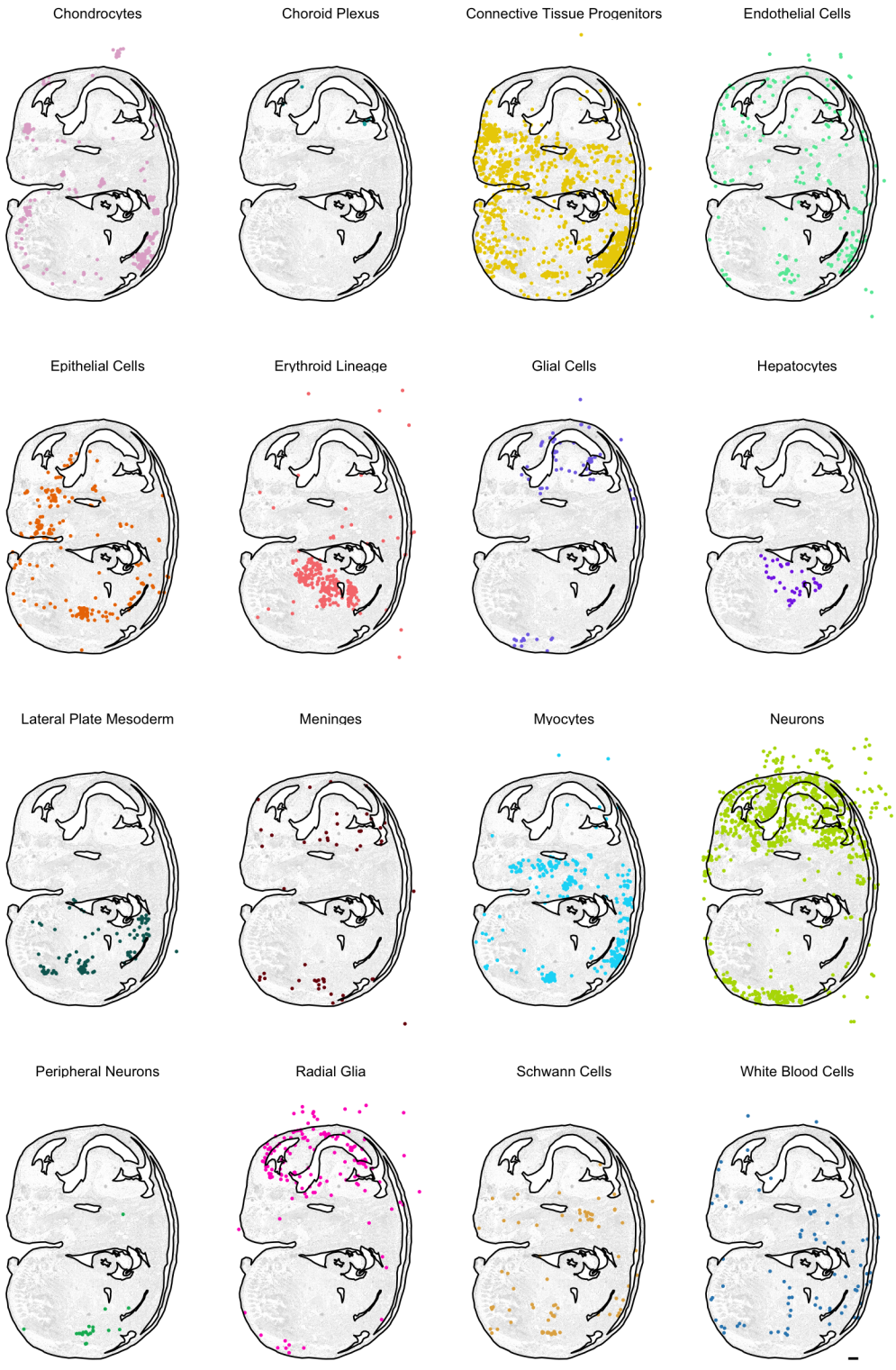

Figure 3.33: **Spatial position of cell types for Slide 8 from embryo 1.** Sequenced cells, broken out by cell type, are placed onto the imaged embryo at the spatial position to which they map on top. Top 16 abundant cell types are shown with a 500um scale bar (bottom right).

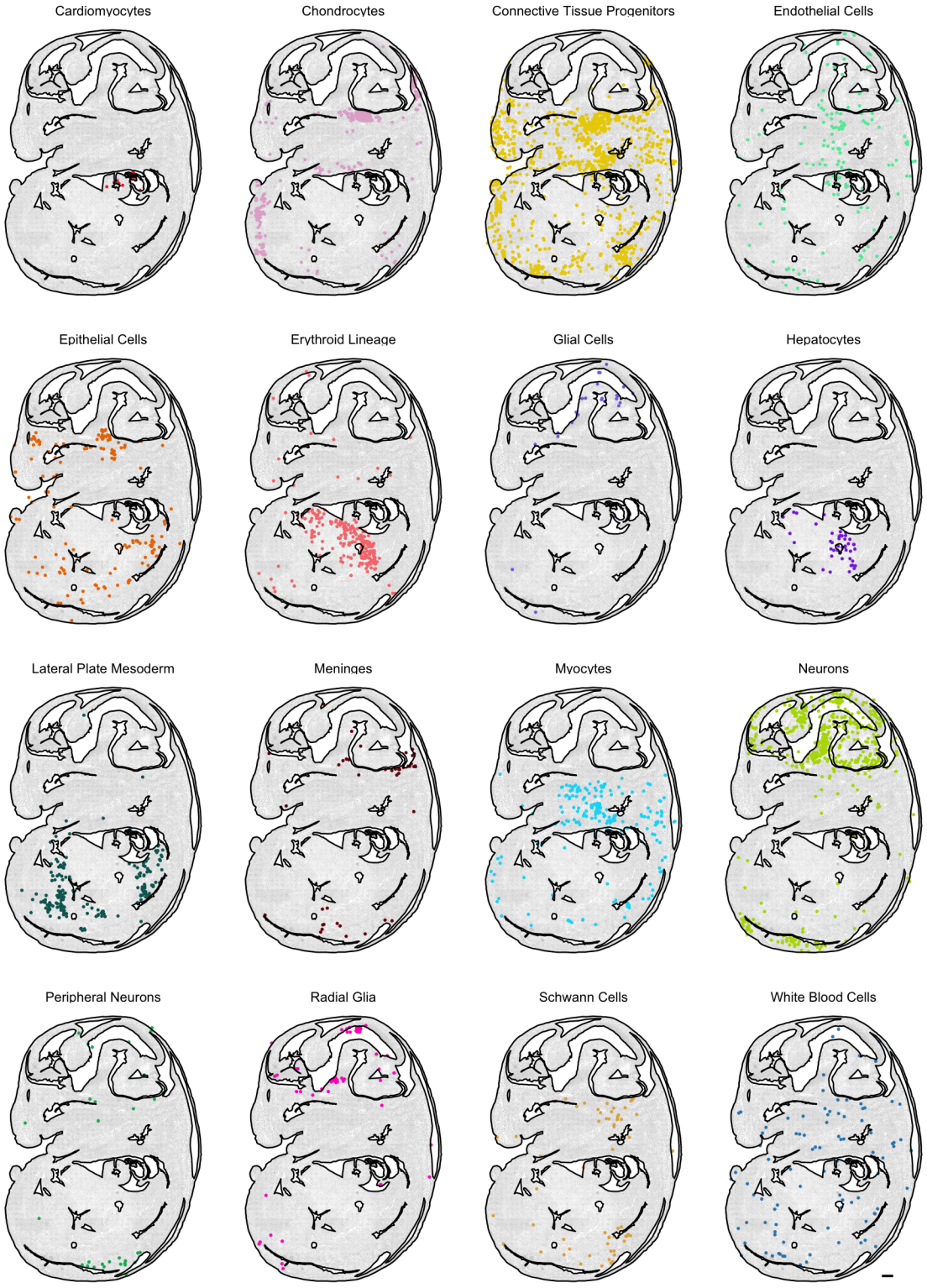

Figure 3.34: **Spatial position of cell types for Slide 8 from embryo 1.** Sequenced cells, broken out by cell type, are placed onto the imaged embryo at the spatial position to which they map on top. Top 16 abundant cell types are shown with a 500um scale bar (bottom right).

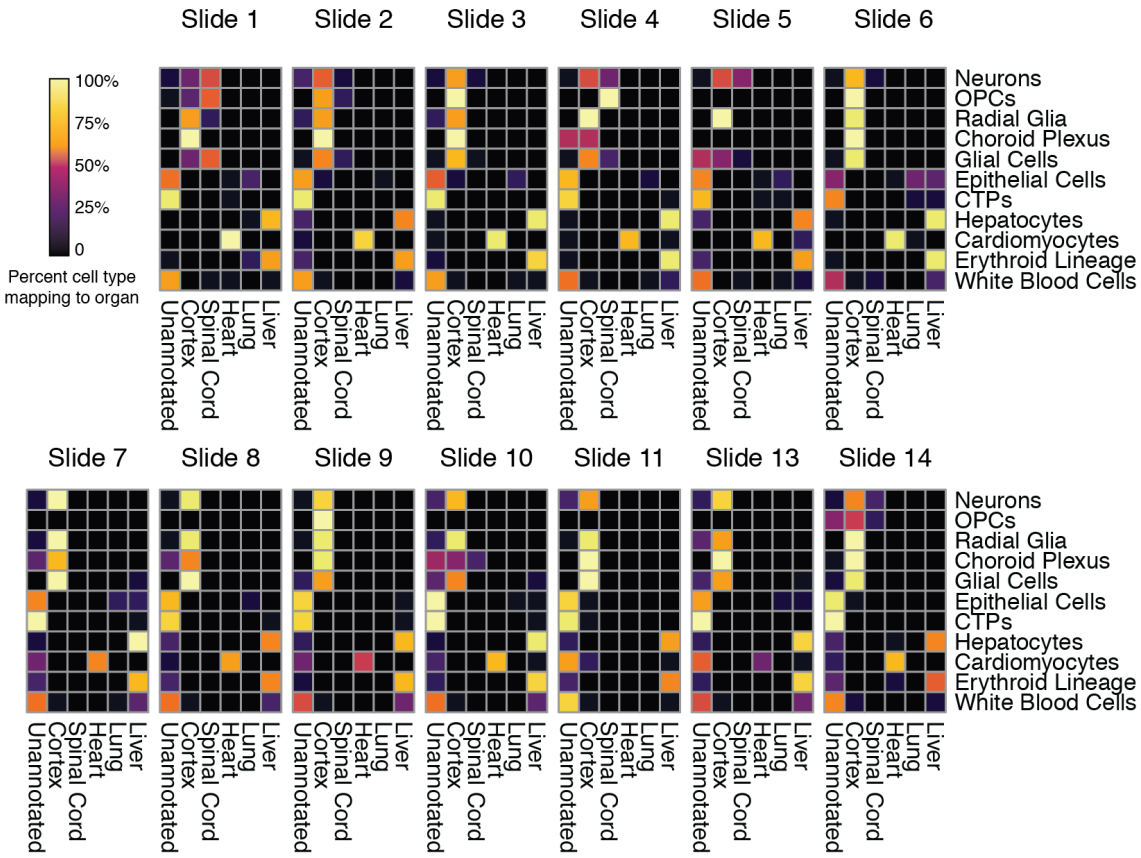

Figure 3.35: **Cell types enriched in annotated anatomical segments.** Heatmap displaying the percentage of cells of a given type mapping to an anatomical annotation. Unannotated denotes regions of the embryo which were not assigned to an anatomical annotation. (OPCs - Oligodendrocyte Progenitor Cells; CTPs - Connective Tissue Progenitors).

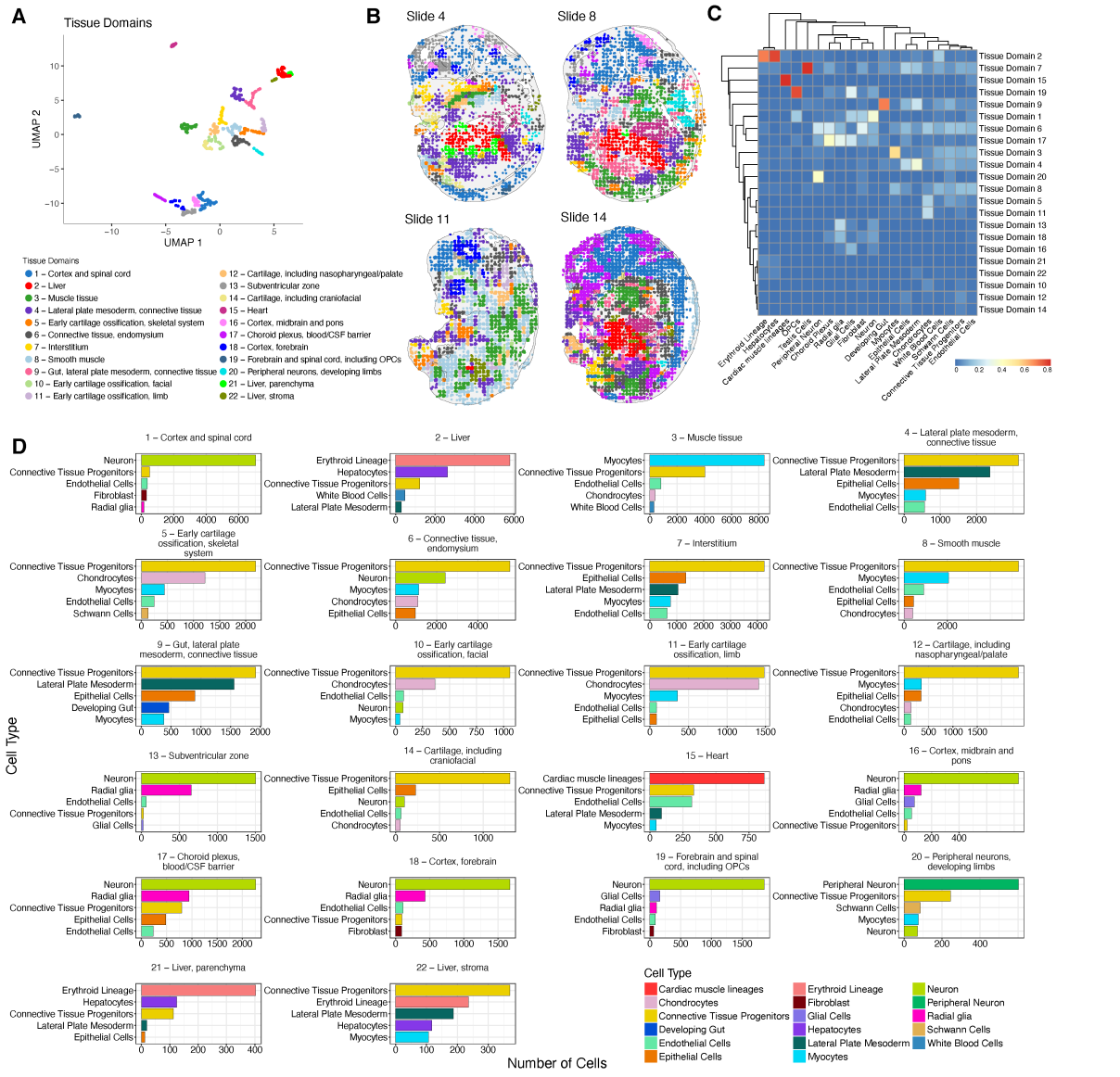

Figure 3.36: **Tissue domains based on similar cell type compositions are found across the embryo. (A)** UMAP dimensionality reduction displaying the tissue domains, identified using the hidden Markov random field model from the Giotto (*108*) package, from each slide based on similar cell type composition in spatial locations. Each point is a single domain from a single slide and is colored according to cluster assignment. **(B)** Cells are colored based on membership to a tissue domain in four representative slides. **(C)** Heatmap displaying the percentage of a cell-type that composes each tissue domain. **(D)** The top five cell types that compose each tissue domain are shown with absolute counts.

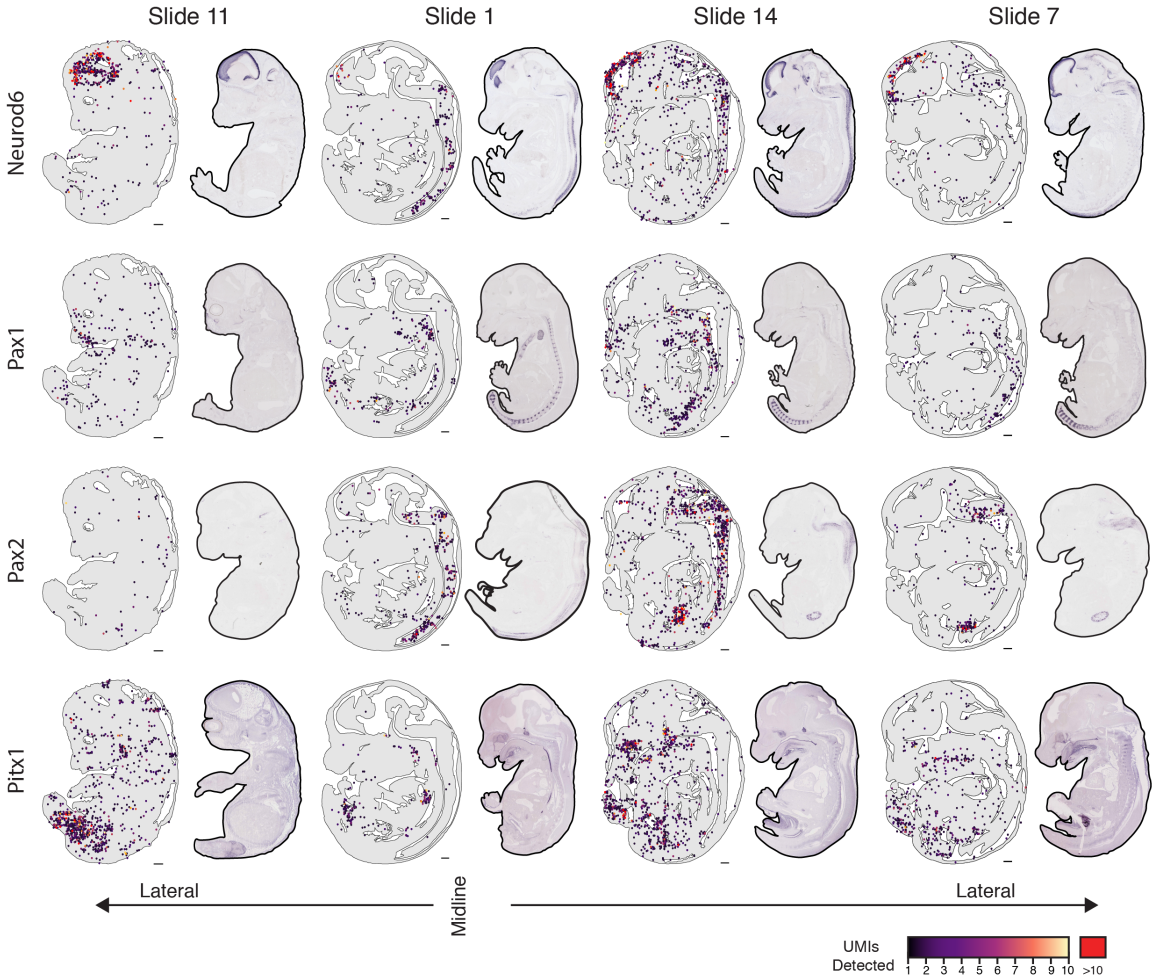

Figure 3.37: **Comparison of traditional ISH and digital in-situs.** Expression for a given gene (rows) plotted for slides from embryo 1 (slides 11,1,14 and 8) (columns) compared to published in-situ hybridization data (*109*) on matched sections. Genes with high spatial autocorrelation in different tissues were chosen for display. 500um scale bar shown at the bottom right of sci-Space data.

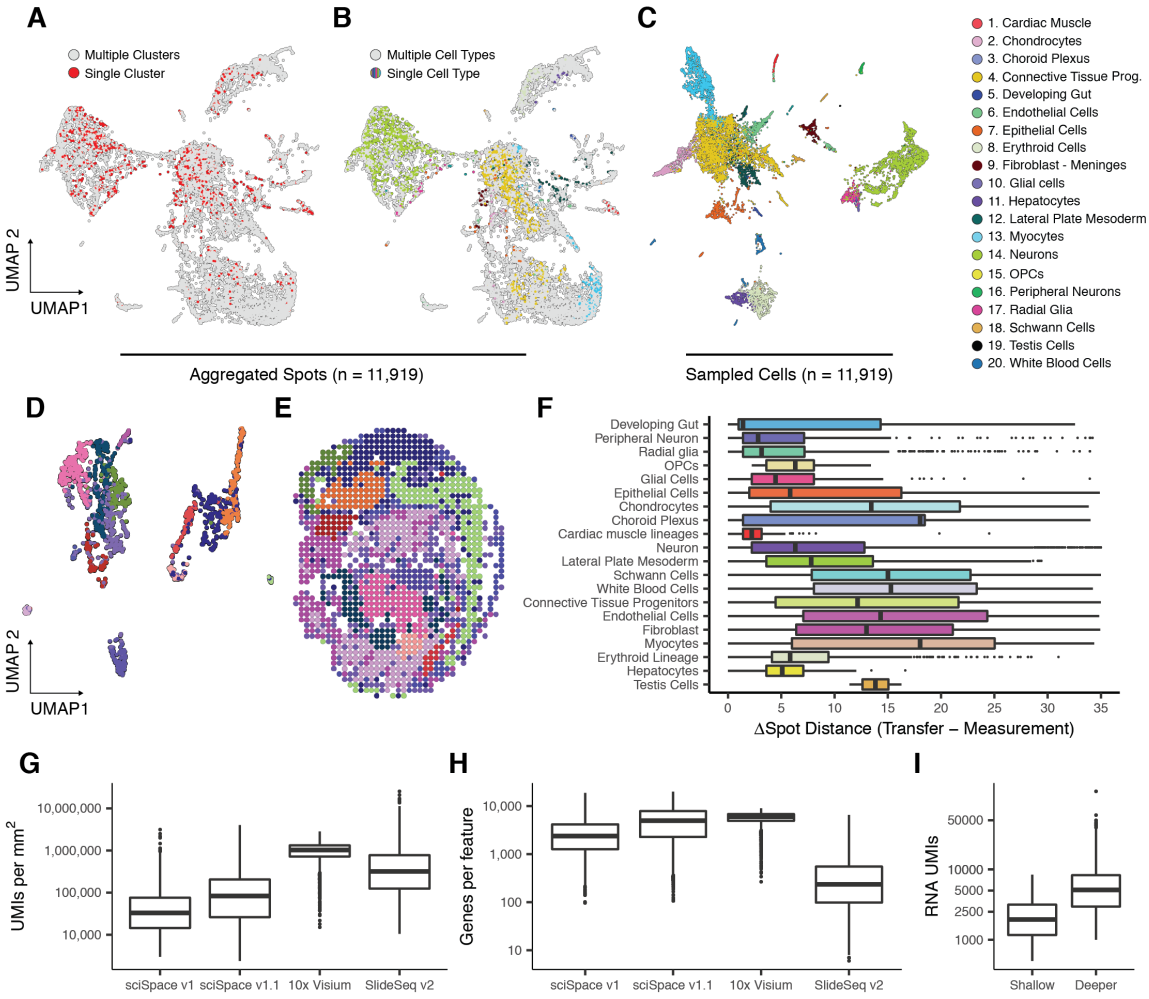

Figure 3.38: **Comparison spatial single cell data and aggregated spatial data.** **(A,B)**, UMAP dimensionality reduction of cells aggregated by spatial positions from all slides. Aggregated spatial positions which contain cells originally mapping to a single UMAP cluster are colored red in panel **(A)**, and spatial positions consisting of a single cell type label are colored by their corresponding label in panel **(B)**. Positions containing cells with membership with multiple clusters or cell types are colored in grey. **(C)** UMAP dimensionality reduction of single cells sampled from each spatial position, colored by cell type label. **(D,E)** Seurat v3 aggregated spatial positions from Slide 14 only. **(D)** displayed in UMAP space and colored by cluster or **(E)** displayed in their spatial positions on the embryo. **(F)** Distance between the Seurat v3 highest probability mapping of a single cell onto the aggregated spatial positions and a cell's recovered position by sci-Space. Distance calculated between positions on the spatial grid. **(G,H)** Comparison between sci-Space data (v1 refers to slides 1-6 and v1.1 refers to slides 7-14), 10x Visium (mouse cortex) data and Slide-seq-v2 (hippocampus) data displaying UMIs/mm2 in panel **(G)** or genes detected per spatial feature in panel **(H)**. **(I)** The current distribution of UMIs/cell of sci-Space data (4,831 reads/cell) versus deeper sequencing (43,536 reads/cell) performed on a small scale sci-Space experiment.

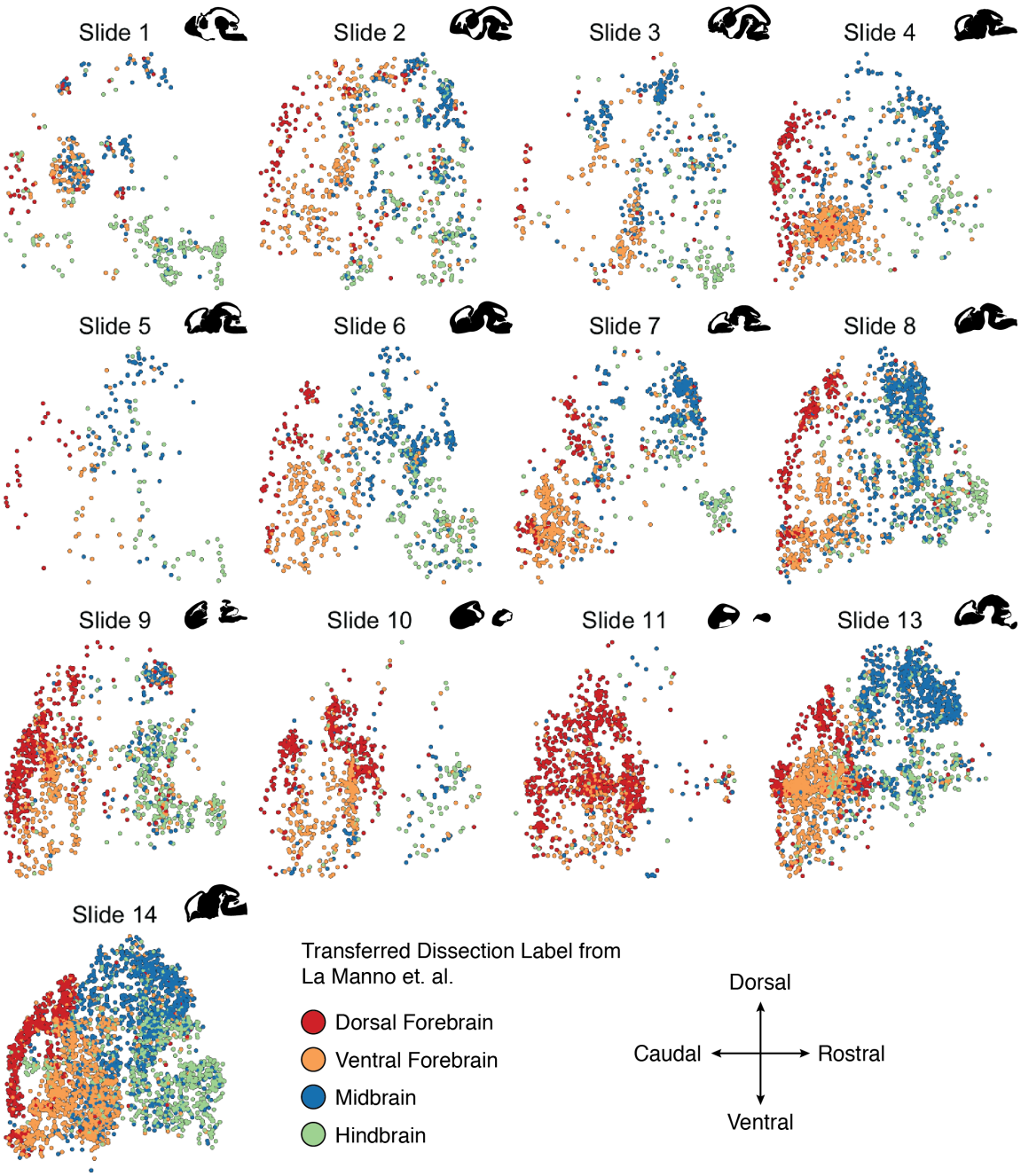

Figure 3.39: **Dissection labels transferred from a developing mouse brain atlas to cells in the sci-Space dataset.** The sci-Space dataset was co-embedded with the developing mouse brain atlas dataset from La Manno et al (*103*). The anatomical dissection label was then transferred from the developing mouse brain atlas dataset to the sci-Space dataset. This was done using the majority label of each cell's five nearest neighbors in the co-embedded space. Each facet shows a different slide with a contour of the brain shown to the right of each facet.

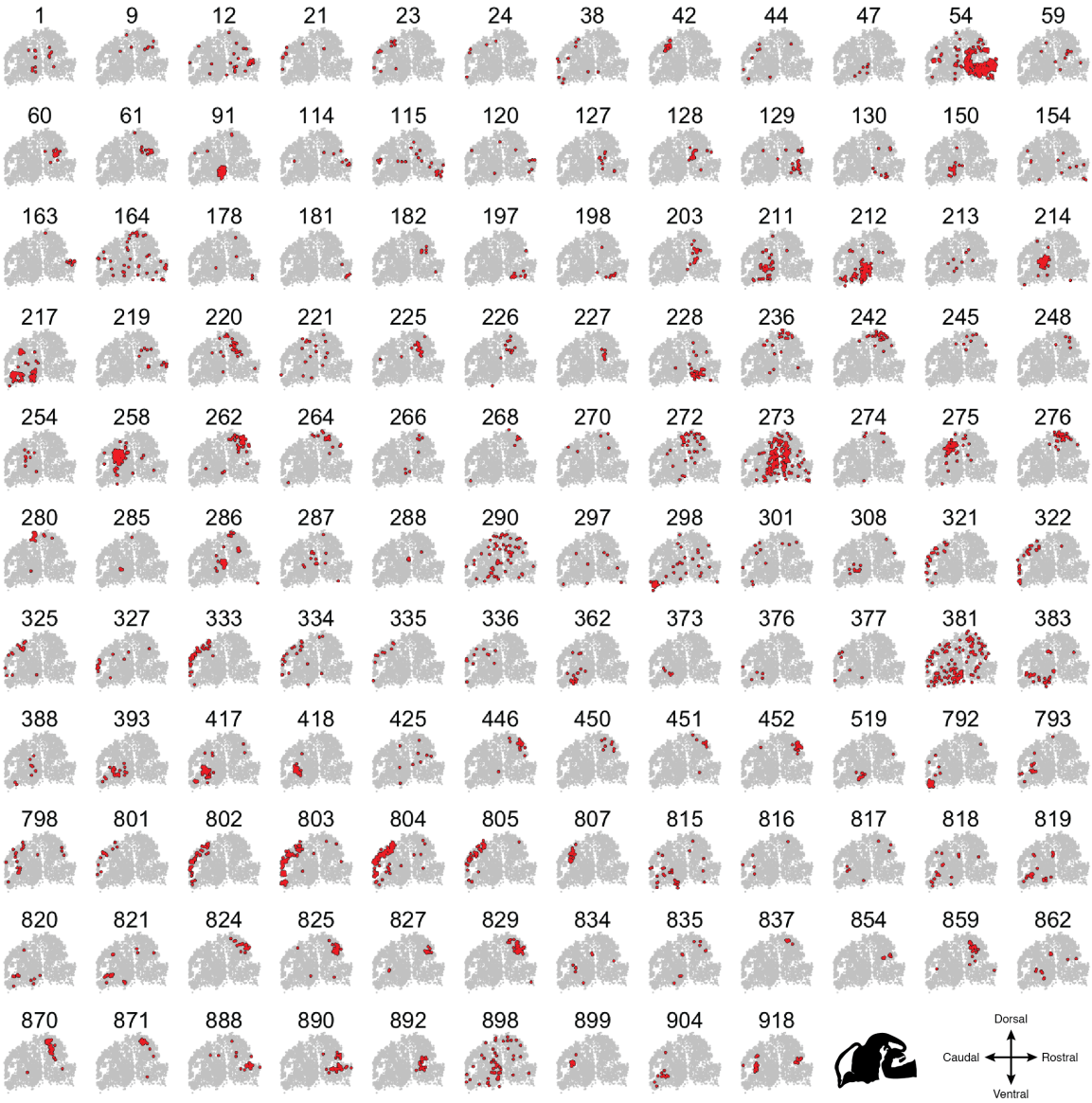

Figure 3.40: **Transfer of spatial labels from a developing mouse brain atlas to the sci-Space dataset – Slide 14.** The sci-Space dataset was co-embedded with the developing mouse brain atlas dataset from La Manno et al. (*103*). The highly resolved UMAP cluster label was then transferred from the developing mouse brain atlas dataset to the sci-Space dataset. This was done using the majority label of each cell's five nearest neighbors in the co-embedded space. Each facet displays all sci-Space cells (in grey) with highlighted cells (in red) bearing the transferred UMAP cluster. Only clusters with greater than 5 cells are shown.

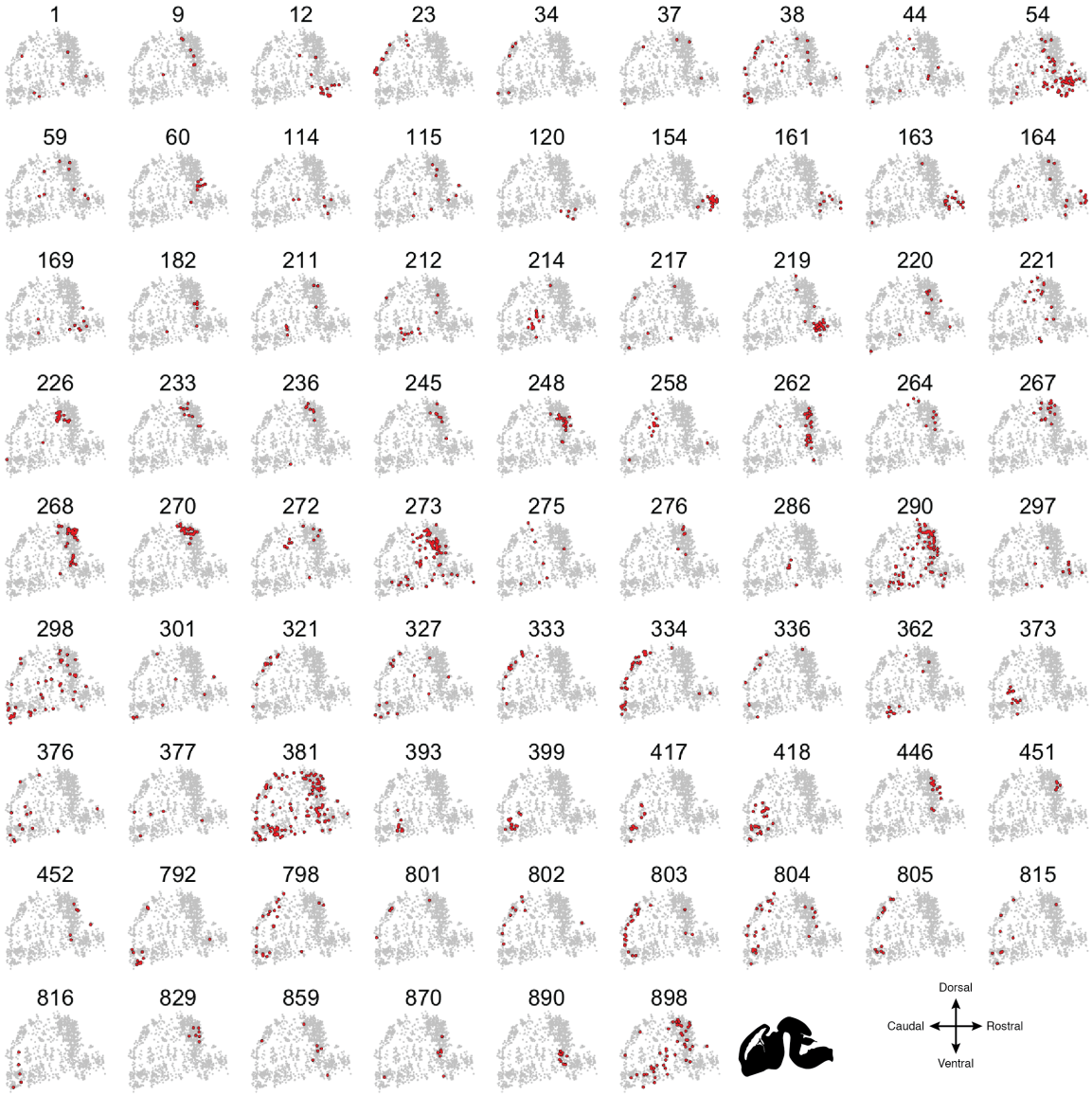

Figure 3.41: **Transfer of spatial labels from a developing mouse brain atlas to the sci-Space dataset – Slide 13.** The sci-Space dataset was co-embedded with the developing mouse brain atlas dataset from La Manno et al. (*103*). The highly resolved UMAP cluster label was then transferred from the developing mouse brain atlas dataset to the sci-Space dataset. This was done using the majority label of each cell's five nearest neighbors in the co-embedded space. Each facet displays all sci-Space cells (in grey) with highlighted cells (in red) bearing the transferred UMAP cluster. Only clusters with greater than 5 cells are shown.

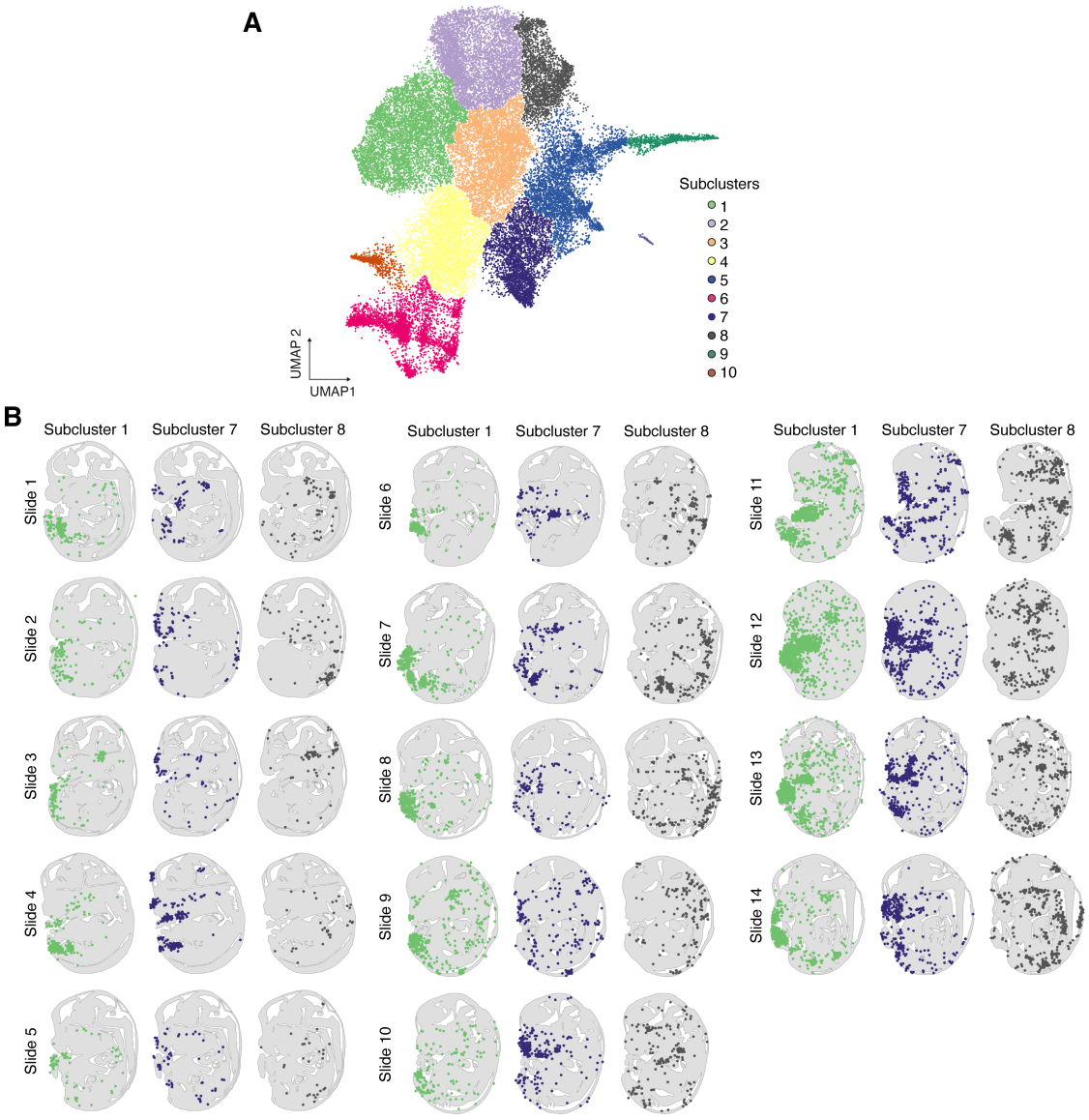

Figure 3.42: **Connective tissue progenitors are composed of multiple subtypes with distinct spatial distributions. (A)** Sub-clustering of the differentiating mesenchyme cells with cells colored by cluster. **(B)** Position of 3 selected sub-clusters that show differing spatial restriction within the embryos, with subcluster 1 primarily in the limbs, subcluster 7 focused in the face, and subcluster 8 more dispersed.

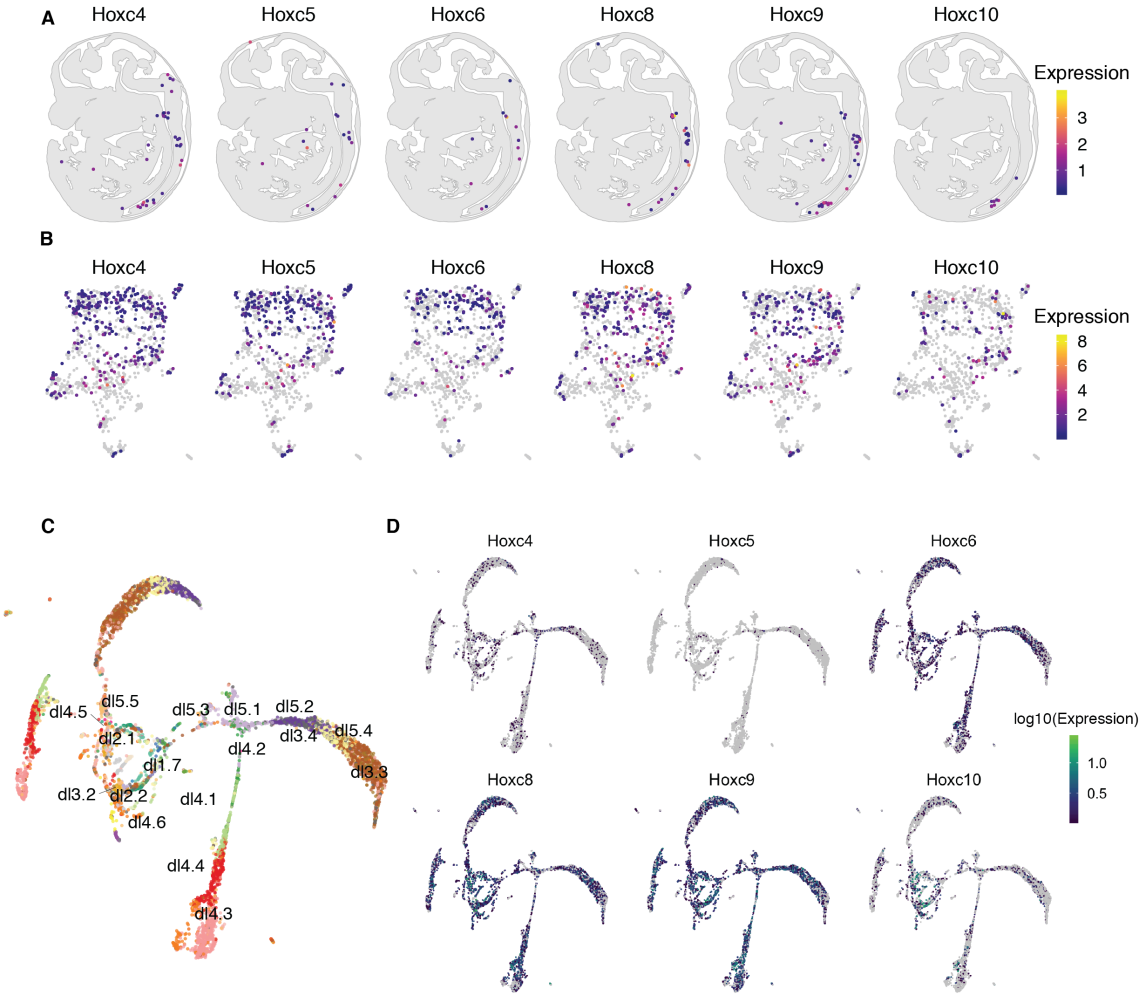

Figure 3.43: **Connective tissue progenitors are composed of multiple subtypes with distinct spatial distributions.** Subtypes of the neurons from Slide 1 annotated using developing mouse brain and spinal cord atlases (*103*, *124*) plotted **(A)** by UMAP and **(B)** spatially. **(C)** log10-scale boxplot of the UMAP Moran's I statistic for Hox genes displayed in **(Figure 3.3B)** versus all other expression level-matched genes (p-value < 0.01, two sided t-test). **(D)** UMAP embedded expression patterns for HoxA genes. (E) UMAP embedding of neurons from the E13.5 stage of mouse development from a published single cell spinal cord dataset (*124*). Colors and labels mark neuron subtypes annotated by the authors. **(F)** Expression of the same HoxA cluster displayed in panels (A-B).

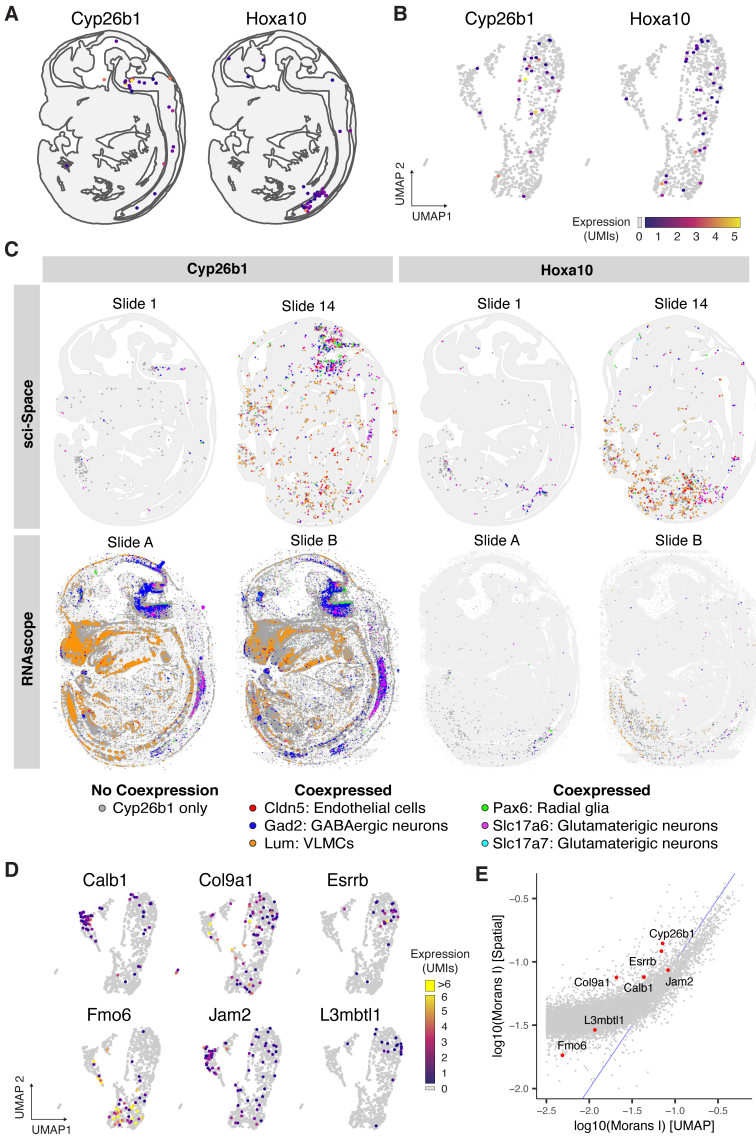

Figure 3.44: **Connective tissue progenitors are composed of multiple subtypes with distinct spatial distributions.** **(A)** Spatial patterns of *Cyp26b1* and *Hoxa10* gene expression for neurons from Slide 1 and **(B)** corresponding expression patterns across neuronal cell states/subtypes (UMAP embedded). **(C)** Cyp26b1 and marker gene coexpression patterns measured by sci-Space (top row) were validated by comparison with co-expression detected with RNA FISH (bottom row). Cell types were designated using the following marker genes: *Cldn5* - endothelial cells, *Gad2* - GABAergic (inhibitory) neurons, Lum - vascular and leptomeningeal cells (VLMCs), *Pax6* - radial glia, *Slc17a6* and *Slc17a7* - Glutamatergic (excitatory) neurons. RNA FISH supported the spatial mapping of sci-Space transcriptomes expressing *Cyp26b1* and *Hoxa10* and the assayed marker genes. **(D)** Expression patterns (UMAP) for a subset of spatially but not cell state restricted gene across neuronal cell states (Slide 1). **(E)** Analysis of Slide 14 shows consistent spatial versus UMAP log10(Morans I) values as compared to Slide 1 **(Figure 3.3E)**.

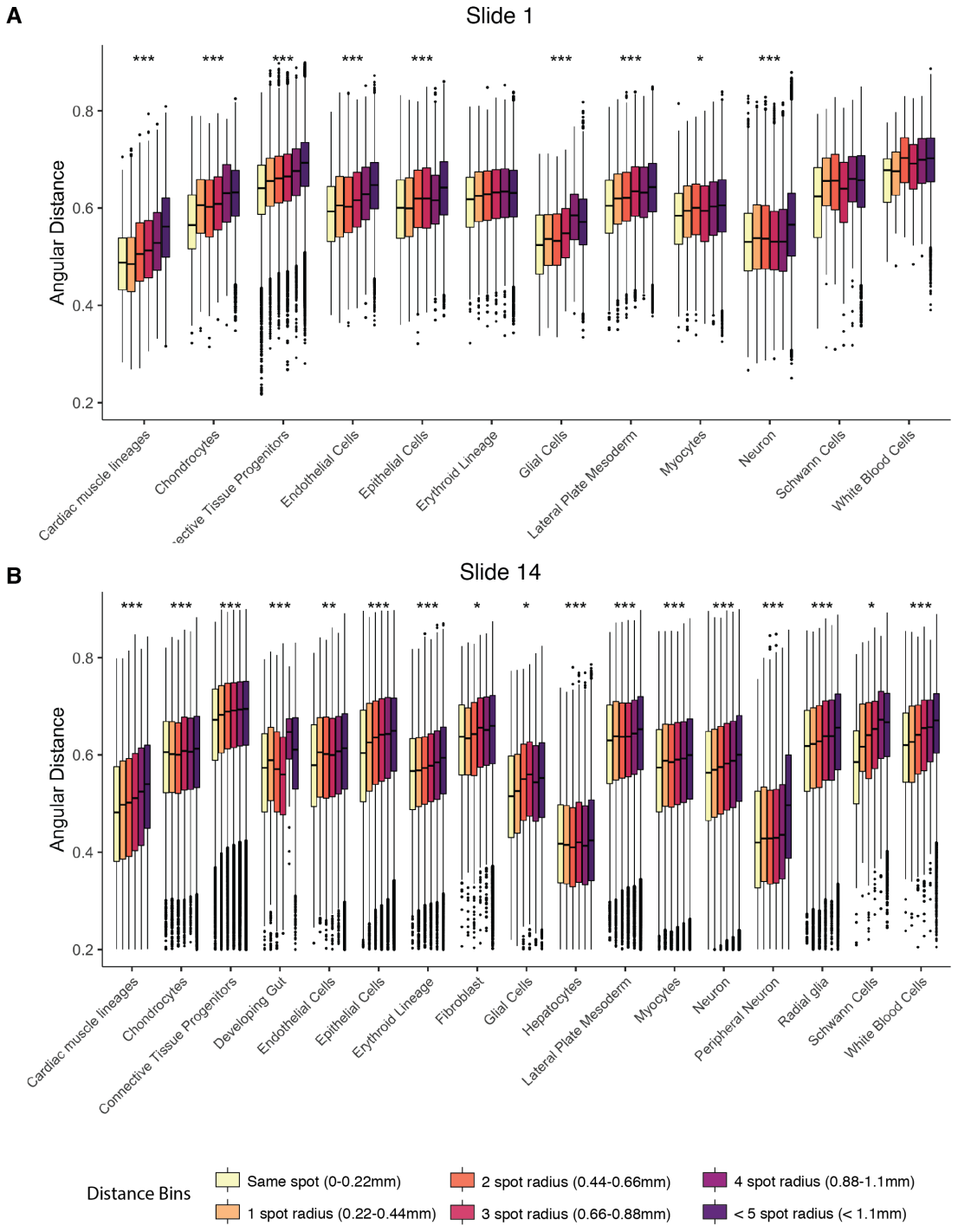

Figure 3.45: **Similarity between pairs of transcriptomes as a function of spatial distance. (A,B)** Pairwise angular distance between cells broken out by the distance in millimeters between cells mapping back to the spatial grid. Boxplots display all cell types with over 100 cells in **(A)** Slide 1 and **(B)** Slide 14. Stars denote significant distance coefficient in linear regression without the application of an estimate filtering step (* : p-value < 0.01, ** : p-value < 0.001, *** : p-value < 0.0001, Wald linear regression test; Methods).

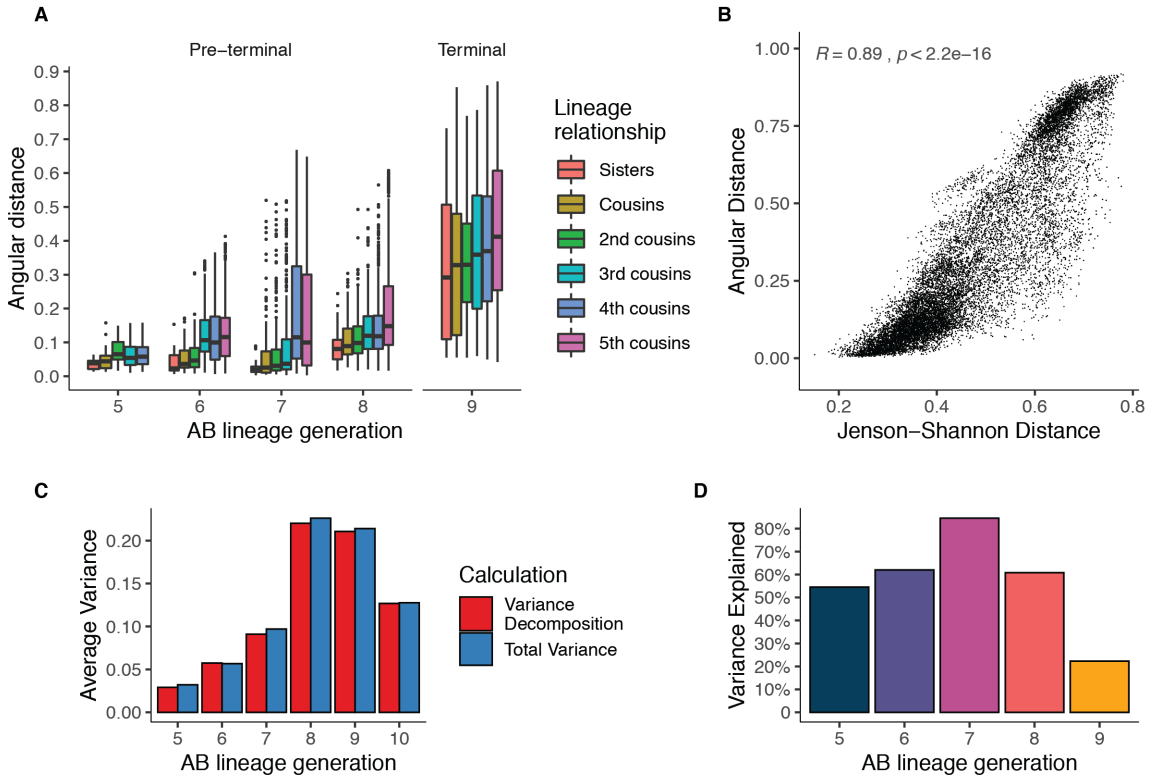

Figure 3.46: **Re-examining the effect of lineage on transcriptome in the developing C. elegans dataset using angular distance.** **(A)** Pairwise angular distance between cells in the C. elegans AB lineage (*85*). Boxplots denote lineage relationship between the pair of cells. **(B)** Relationship between Jenson-Shannon distance and angular distance calculated between all pairs of cells displayed in panel **(A)**. **(C)** Empirical confirmation of the Law of Total Variance. Red bars indicate the sum of unexplained and explained variance after grouping cells that shared a common parent. Blue bar denotes the average mean squared error from the global mean. **(D)** Variance explained for cells in the AB lineage after accounting for parental identity. Variance explained statistic was computed as described in Methods.

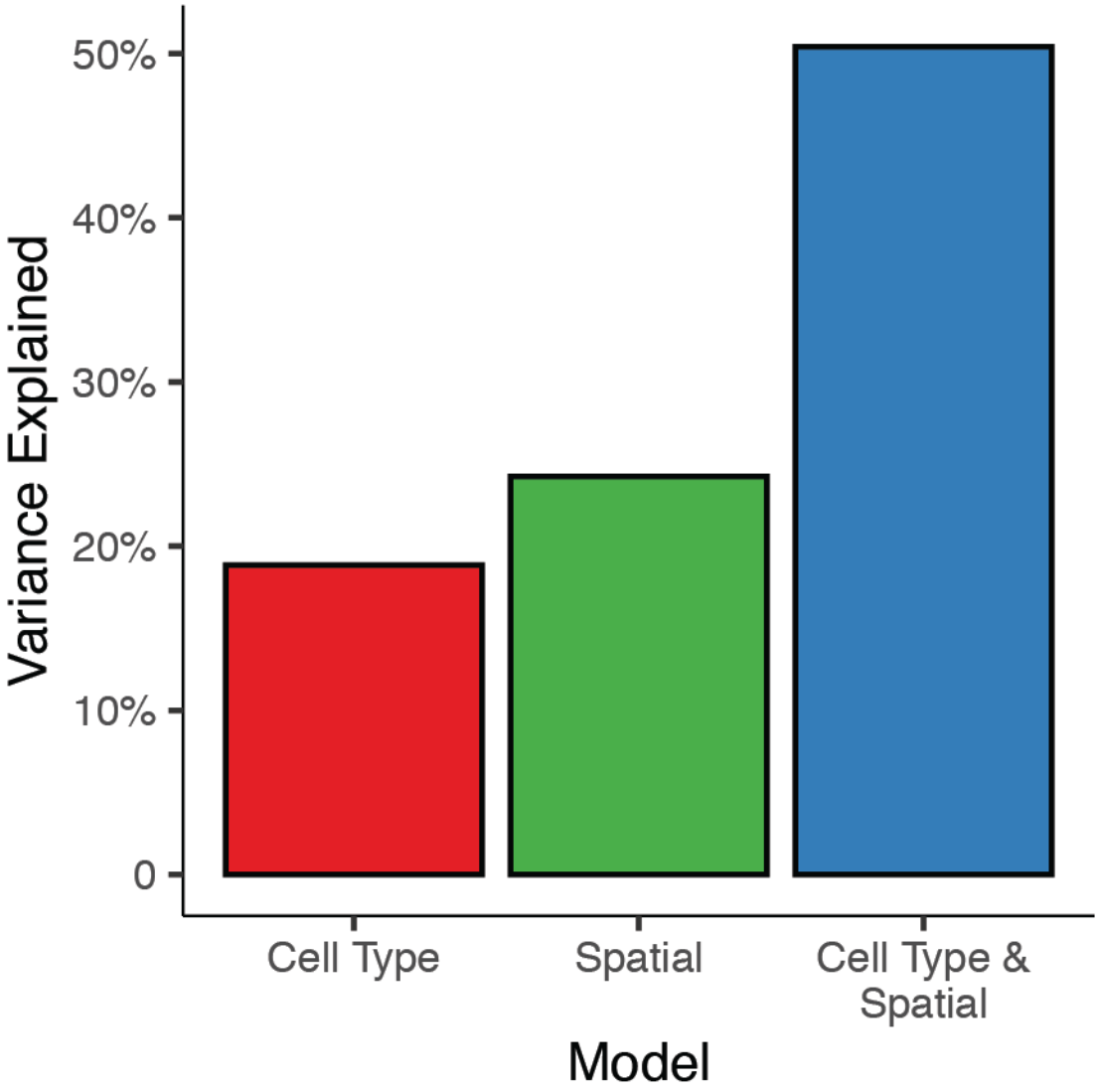

Figure 3.47: **Variance explained by the cell type label, spatial label, or both.** Proportion of variance explained, apart from that attributable to sparse UMI sampling, as described in Methods. Each spatial bin corresponds to 4 adjacent spots that are collapsed. Spatial bins that only contained a single cell of a given cell type were excluded from the analysis.

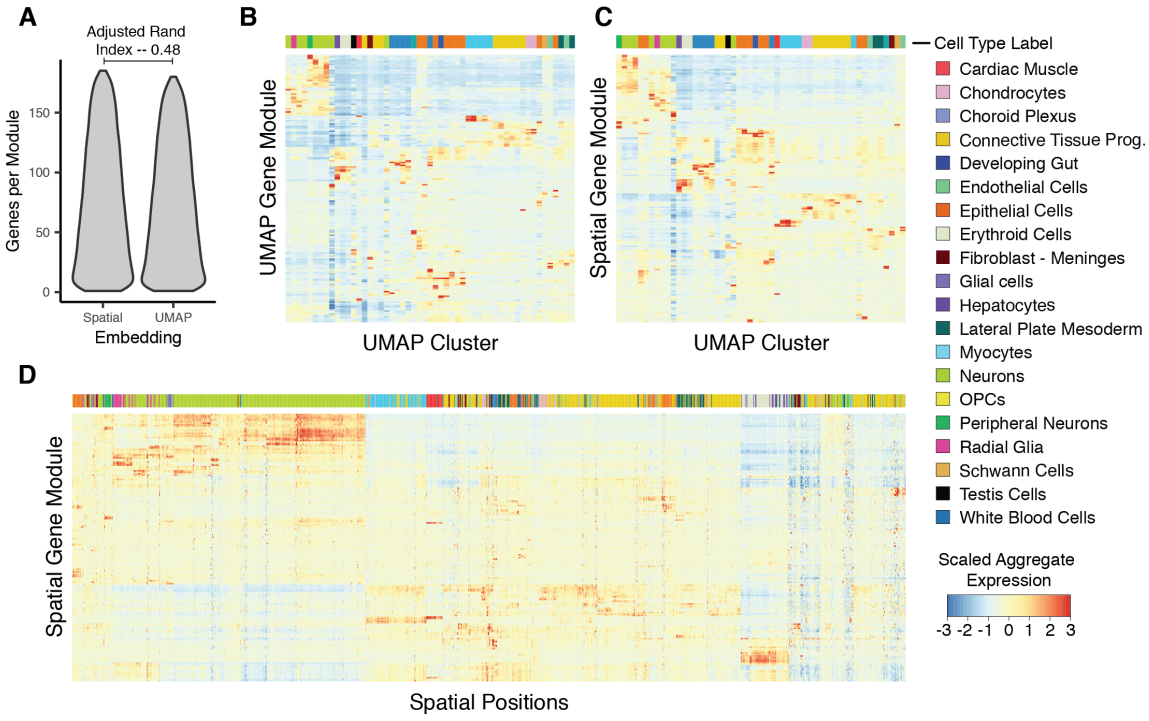

Figure 3.48: **Variance explained by the cell type label, spatial label, or both.** **(A)**Gene modules were recovered by using either the aggregated spot by gene count matrix (Spatial) or the cell-by-gene count matrix (UMAP). Panel A shows the distribution of genes per module and the adjusted rand index calculated on the two groupings. **(B,C)** Row- and column-clustered heatmap of aggregated **(B)** UMAP-derived module expression (rows) per UMAP cluster (columns) or **(C)** spatially-derived module expression (rows) per UMAP cluster (columns). Color bar on top corresponds to the majority cell type within the UMAP cluster. **(D)** Row- and column-clustered heatmap of aggregated, spatially-derived module expression (rows) per spatial position (columns). Color bar on top corresponds to the majority cell type within each spatial position.

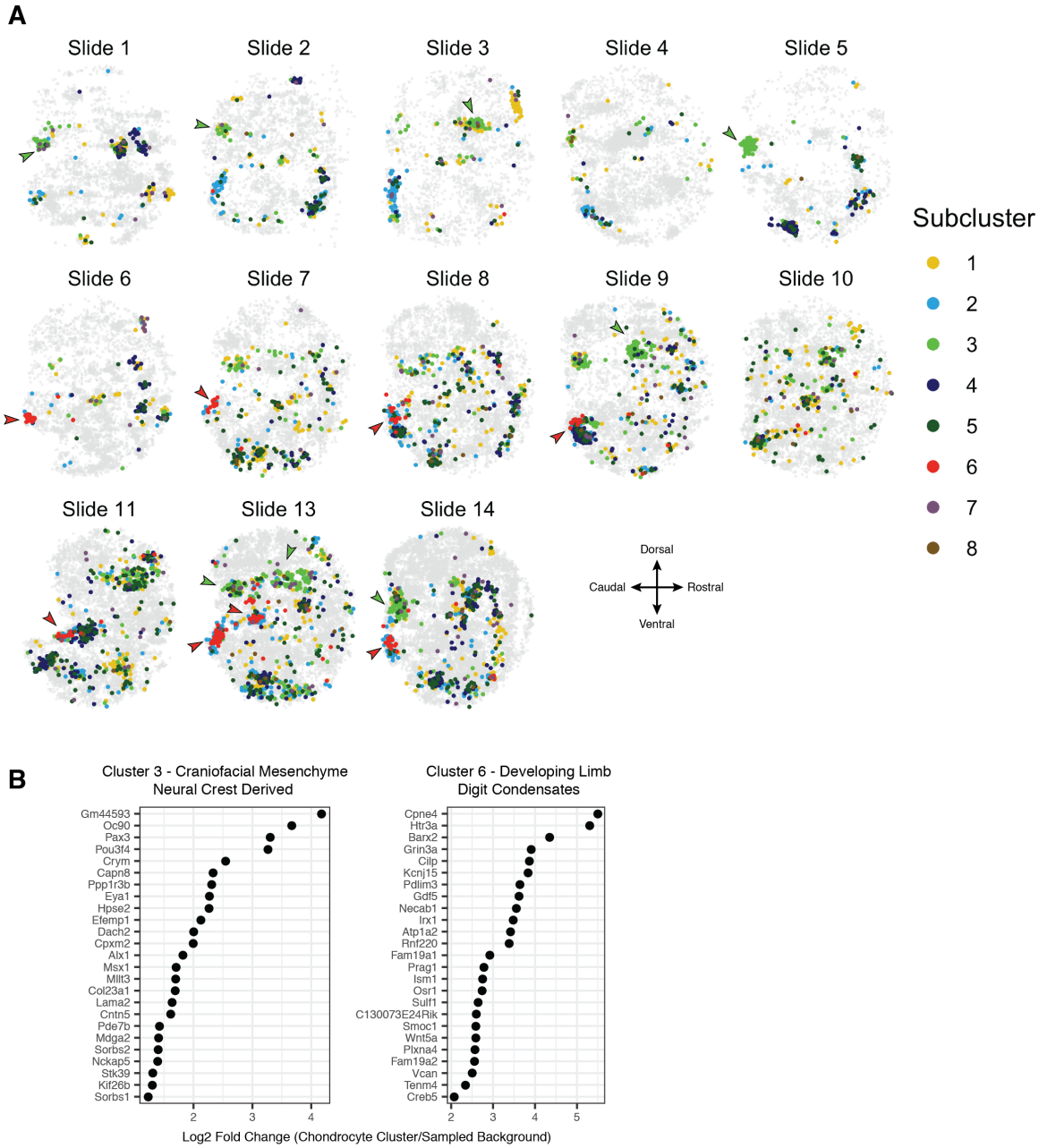

Figure 3.49: **Subtypes of chondrocytes are spatially restricted.** **(A)** Sub-clustered chondrocytes colored by cluster. Colored arrowheads highlight accumulations of cluster 6 (blue) and cluster 7 (purple). **(B)** Log2-fold change of genes expressed in cluster 3 (left) and cluster 6 (right) relative to a background distribution of sampled chondrocyte cells.

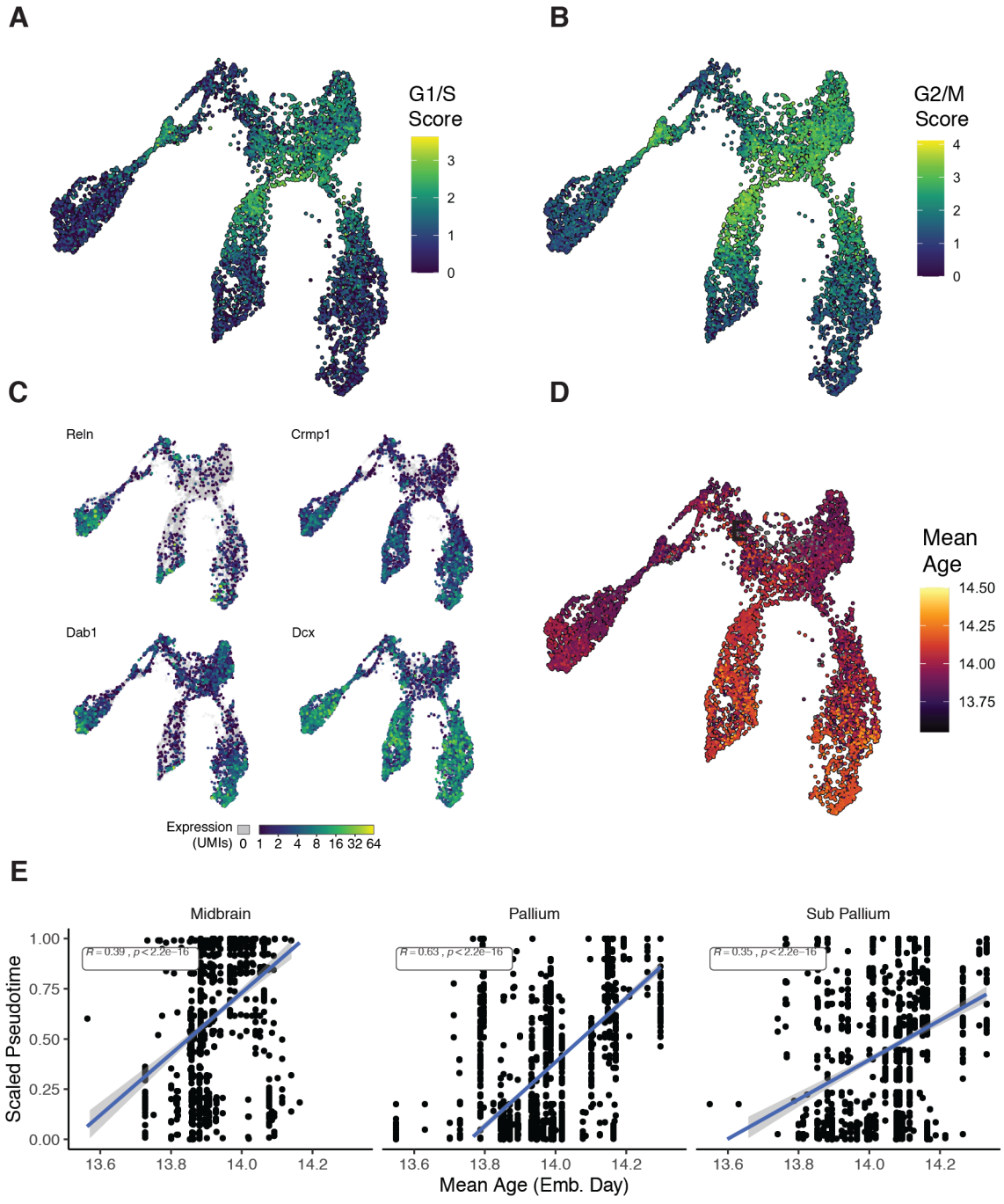

Figure 3.50: **Characteristics of neural pseudotemporal trajectory.** **(A-D)** UMAP embedding of neural cells in the trajectory colored by **(A)** aggregate expression of G1/S markers, **(B)** aggregate expression of G2/M markers, (**C**) selected gene expression with documented roles in neuronal migration, or (**D**) age of transferred nearest neighbors in the developing brain atlas dataset (*103*). **(E)** Scatter plot displaying transferred age and pseudotime for each of the three trajectories with a least squares regression fit (blue).

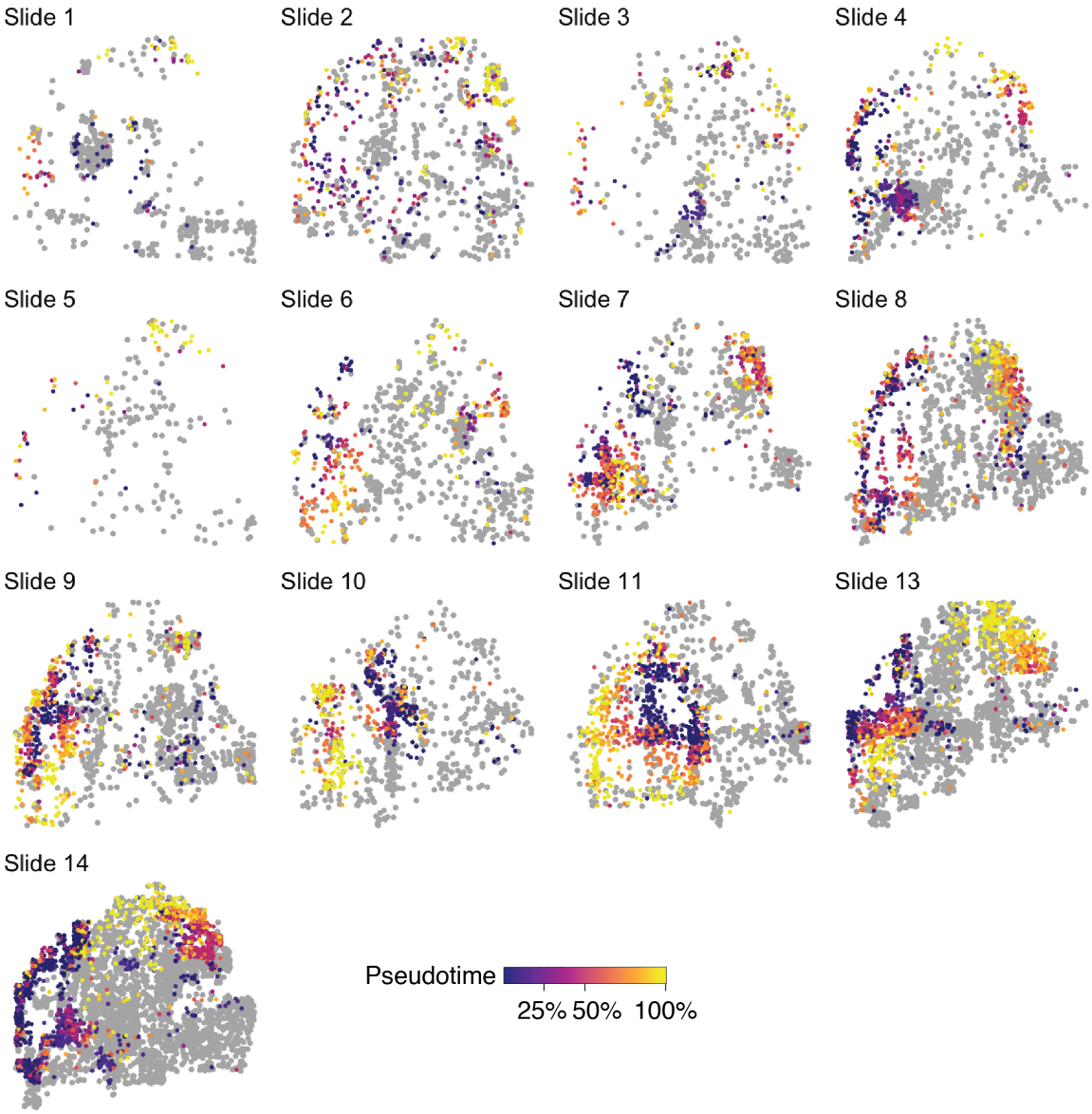

Figure 3.51: **Spatial location of cortical neural trajectories.** Neurons and radial glia identified in the neural trajectories are shown colored from navy blue (early in pseudotime) to yellow (late in pseudotime), while neurons and radial glia not found in the trajectory are colored grey.

‘

Chapter 4: CONCLUSION

4.1 NEW MODELS FOR A NEW ERA

The sequencing of the human genome led to a sea change in the way that we perform science. Another sea change, which has yet to realize its potential, is the training of highly flexible computational models for generalized problems in biology. Epitomized by the training of the AlphaFold algorithm(*125*), computational models are beginning to perform complex tasks with startling accuracy. In the case of AlphaFold, the development of new model architectures, along with the shedding of old heuristics, has led to the development of an astonishingly accurate algorithm for predicting protein structure. Although Alphafold currently performs the task of protein structure prediction, there is a clear opportunity for such an algorithm to perform de novo protein design – the task of designing a sequence given a target structure. In another avenue, the success of the language model GPT3, is predicated on learning 175 billion model parameters by training on massive amounts of diverse training data. The result is an algorithm that can autocomplete text, answer questions, and generate language de novo capable of passing the Turing test.

In human genetics efforts to achieve a similar scale are currently in progress. To date, the Exome Aggregation Consortium (ExAC) database and the Genome Aggregregation Database (GnomAD) (*126*) have amassed a total of 125,748 exome sequences and 15,708 whole genome sequences. Importantly, this data is harmonized and filtered for high quality sequencing data allowing for its use in future efforts. Furthemore, the incorporation of metadata pertinent to the sequenced individual(*127*) has led to the development of data driven models that link disease observed in the human population to the underlying genotypic variants. Comparable efforts are currently underway to sequence diverse human populations which will only add to the value of this resource(*128*).

4.2 HARMONIZING SINGLE CELL GENOMICS

Analogous to the commercial activity that led to the proliferation of next generation sequencing, the commercialization of single cell genomics platforms has led to their widespread adoption. This means that the amount of high quality, single cell data produced will also increase exponentially as the price of single cell library preparation and sequencing drop. Accordingly, several groups have already invested in the development of software for the integration of different single cell datasets(*129*, *130*). These algorithms operate by matching similar cells found in the two datasets to derive a new projection. The result of this procedure means that any integration between two unique sets of data, is itself unique.

To solve this problem there have been attempts to produce a reference atlas of cells. The idea is that when a single cell experiment is performed, cells can be mapped onto this reference, much like the mapping of DNA fragments to the genome. This procedure, which was recently deployed on circulating peripheral blood mononuclear cells(*131*), was aided by deep characterization of the cells in the reference atlas. In this study, the authors measured surface protein expression in conjunction with the cell's transcriptome to produce a cellular atlas resolved by both modalities. Notably, measurement of surface marker expression in PBMCs(*132*) allows for the integration of a deep knowledge base, which has been developed by immunologists over decades. The result is a dataset that not only allows for computational integration, but also integration with scientists in the field of immunology. This synergy will surely help harmonize existing immunology paradigms with the new, high resolution measurements that are being made currently.

4.3 STRUCTURE-FUNCTION

The key insights from Hao et. al's attempt to define a reference is twofold: 1) The field needs a stable reference to facilitate comparisons and consistent annotations, and 2) this reference needs to be readily integratable into existing systems of knowledge. However, the chosen measurement set (surface protein expression and gene expression) may be uniquely suited to PBMCs. Unlike cells in solid tissues and organs, PBMCs do not have a definitive spatial context, thus the omission

of spatial context is not an issue. However, for the mapping of cells from non-liquid tissues and organs, the capture of spatial information will be necessary.

To address this problem, I believe that the field must start a collective effort to construct a set of single cell transcriptomic atlases for each organism that are both spatially- and lineally-resolved. This set of atlases spanning the stages of development, adolescence, and adulthood, would effectively provide a set of reference "structures" for each organism. In this structural representation each cell would be akin to an atom in an atomic structure and different cell types would stand in for different atom types. The connectivity between cells could either be represented as the spatial connectivity between cells or the lineal relationships between them.

The development of such a resource, modeled on the genome projects, could then act as the reference organismal atlas. Cells sequenced at any future time could then get mapped to this reference atlas. Furthermore, this representation of an organism could act as a cipher's key, unifying sequencing based readouts and imaging based readouts. This includes opportunities for training machine learning models that convert from another imaging modality (such as histology or radiographic imaging) to single cell data, and vice-versa.

Finally, such a representation would be useful in training models like AlphaFold. Instead of learning the structures of proteins, a model could be trained to learn the connectivities between cells that form tissues and how tissues form organs. Analogous to the fold in protein structure, spatial motifs of cell types could be identified and established at the cellular and molecular levels. Truly fantastic extensions of such a model would enable the de novo design of new tissue architectures, and eventually new organisms. However, to achieve this future, the training data needed to train such a model would need to be diverse – spanning the tree of life and various pathological states.

After solving the structure of the organism, the number of training examples needed to train a useful model will require new measurement technologies. These technologies will need to be applicable across many tissue contexts, fast and highly multiplexed. Although the technologies presented in this work begin to achieve these characteristics, there are many refinements still needed to advance this lofty vision.

BIBLIOGRAPHY

(*1*) Erika Check Hayden. "Technology: The $1,000 genome". en. In: *Nature* 507.7492 (Mar. 2014), pp. 294–295.

(*2*) E S Lander et al. "Initial sequencing and analysis of the human genome". en. In: *Nature* 409.6822 (Feb. 2001), pp. 860–921.

(*3*) Karen H Miga et al. "Telomere-to-telomere assembly of a complete human X chromosome". en. In: *Nature* 585.7823 (Sept. 2020), pp. 79–84.

(*4*) Glennis A Logsdon et al. "The structure, function and evolution of a complete human chromosome 8". en. In: *Nature* 593.7857 (May 2021), pp. 101–107.

(*5*) Mouse Genome Sequencing Consortium and Mouse Genome Sequencing Consortium. *Initial sequencing and comparative analysis of the mouse genome*. 2002.

(*6*) Chimpanzee Sequencing and Analysis Consortium. "Initial sequence of the chimpanzee genome and comparison with the human genome". en. In: *Nature* 437.7055 (Sept. 2005), pp. 69–87.

(*7*) Kerstin Howe et al. "The zebrafish reference genome sequence and its relationship to the human genome". en. In: *Nature* 496.7446 (Apr. 2013), pp. 498–503.

(*8*) J C Venter. *GENOMICS: Shotgun Sequencing of the Human Genome*. 1998.

(*9*) V E Velculescu et al. "Serial analysis of gene expression". en. In: *Science* 270.5235 (Oct. 1995), pp. 484–487.

(*10*) Paul Bertone et al. "Global identification of human transcribed sequences with genome tiling arrays". en. In: *Science* 306.5705 (Dec. 2004), pp. 2242–2246.

(*11*) Jill Cheng et al. "Transcriptional maps of 10 human chromosomes at 5-nucleotide resolution". en. In: *Science* 308.5725 (May 2005), pp. 1149–1154.

(*12*) David S Johnson et al. "Genome-wide mapping of in vivo protein-DNA interactions". en. In: *Science* 316.5830 (June 2007), pp. 1497–1502.

(*13*) Gregory E Crawford et al. "Genome-wide mapping of DNase hypersensitive sites using massively parallel signature sequencing (MPSS)". en. In: *Genome Res.* 16.1 (Jan. 2006), pp. 123–131.

(*14*) Gregory E Crawford et al. "DNase-chip: a high-resolution method to identify DNase I hypersensitive sites using tiled microarrays". en. In: *Nat. Methods* 3.7 (July 2006), pp. 503–509.

(*15*) Alan P Boyle et al. "High-resolution mapping and characterization of open chromatin across the genome". en. In: *Cell* 132.2 (Jan. 2008), pp. 311–322.

(*16*) Jay Shendure et al. “DNA sequencing at 40: past, present and future”. en. In: *Nature* 550.7676 (Oct. 2017), pp. 345–353.

(*17*) Evan A Boyle, Yang I Li, and Jonathan K Pritchard. “An Expanded View of Complex Traits: From Polygenic to Omnigenic”. en. In: *Cell* 169.7 (June 2017), pp. 1177–1186.

(*18*) Amit V Khera et al. “Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations”. en. In: *Nat. Genet.* 50.9 (Sept. 2018), pp. 1219–1224.

(*19*) Yansheng Liu, Andreas Beyer, and Ruedi Aebersold. “On the Dependency of Cellular Protein Levels on mRNA Abundance”. en. In: *Cell* 165.3 (Apr. 2016), pp. 535–550.

(*20*) Fuchou Tang et al. “mRNA-Seq whole-transcriptome analysis of a single cell”. en. In: *Nat. Methods* 6.5 (May 2009), pp. 377–382.

(*21*) Diego Adhemar Jaitin et al. “Massively parallel single-cell RNA-seq for marker-free decomposition of tissues into cell types”. en. In: *Science* 343.6172 (Feb. 2014), pp. 776–779.

(*22*) Simone Picelli et al. “Smart-seq2 for sensitive full-length transcriptome profiling in single cells”. en. In: *Nat. Methods* 10.11 (Nov. 2013), pp. 1096–1098.

(*23*) D S Tawfik and A D Griffiths. “Man-made cell-like compartments for molecular evolution”. en. In: *Nat. Biotechnol.* 16.7 (July 1998), pp. 652–656.

(*24*) Evan Z Macosko et al. “Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets”. en. In: *Cell* 161.5 (May 2015), pp. 1202–1214.

(*25*) Allon M Klein et al. “Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells”. en. In: *Cell* 161.5 (May 2015), pp. 1187–1201.

(*26*) Darren A Cusanovich et al. “Multiplex single cell profiling of chromatin accessibility by combinatorial cellular indexing”. en. In: *Science* 348.6237 (May 2015), pp. 910–914.

(*27*) Junyue Cao et al. “Comprehensive single-cell transcriptional profiling of a multicellular organism”. en. In: *Science* 357.6352 (Aug. 2017), pp. 661–667.

(*28*) Ryan M Mulqueen et al. “Highly scalable generation of DNA methylation profiles in single cells”. en. In: *Nat. Biotechnol.* 36.5 (June 2018), pp. 428–431.

(*29*) Vijay Ramani et al. “Massively multiplex single-cell Hi-C”. en. In: *Nat. Methods* 14.3 (Mar. 2017), pp. 263–266.

(*30*) Junyue Cao et al. “The single-cell transcriptional landscape of mammalian organogenesis”. en. In: *Nature* 566.7745 (Feb. 2019), pp. 496–502.

(*31*) Marlon Stoeckius et al. “Cell Hashing with barcoded antibodies enables multiplexing and doublet detection for single cell genomics”. en. In: *Genome Biol.* 19.1 (Dec. 2018), p. 224.

(*32*) Jase Gehring et al. “Highly multiplexed single-cell RNA-seq by DNA oligonucleotide tagging of cellular proteins”. en. In: *Nat. Biotechnol.* 38.1 (Jan. 2020), pp. 35–38.

(*33*) Christopher S McGinnis et al. “MULTI-seq: sample multiplexing for single-cell RNA sequencing using lipid-tagged indices”. en. In: *Nat. Methods* 16.7 (July 2019), pp. 619–626.

(*34*) Dylan R Farnsworth, Lauren M Saunders, and Adam C Miller. “A single-cell transcriptome atlas for zebrafish development”. en. In: *Dev. Biol.* 459.2 (Mar. 2020), pp. 100–108.

(*35*) Jonathan S Packer et al. *A lineage-resolved molecular atlas of C. elegans embryogenesis at single-cell resolution*. 2019.

(*36*) Junyue Cao et al. “A human cell atlas of fetal gene expression”. en. In: *Science* 370.6518 (Nov. 2020).

(*37*) James A Briggs et al. *The dynamics of gene expression in vertebrate embryogenesis at single-cell resolution*. 2018.

(*38*) Daniel E Wagner and Allon M Klein. “Lineage tracing meets single-cell omics: opportunities and challenges”. en. In: *Nat. Rev. Genet.* 21.7 (July 2020), pp. 410–427.

(*39*) J R Broach and J Thorner. “High-throughput screening for drug discovery”. en. In: *Nature* 384.6604 Suppl (Nov. 1996), pp. 14–16.

(*40*) D A Pereira and J A Williams. “Origin and evolution of high throughput screening”. In: *Br. J. Pharmacol.* 152.1 (2007), pp. 53–61.

(*41*) David Shum et al. “A high density assay format for the detection of novel cytotoxic agents in large chemical libraries”. en. In: *J. Enzyme Inhib. Med. Chem.* 23.6 (Dec. 2008), pp. 931–945.

(*42*) Channing Yu et al. “High-throughput identification of genotype-specific cancer vulnerabilities in mixtures of barcoded tumor cell lines”. en. In: *Nat. Biotechnol.* 34.4 (Apr. 2016), pp. 419–423.

(*43*) Zachary E Perlman et al. “Multidimensional drug profiling by automated microscopy”. en. In: *Science* 306.5699 (Nov. 2004), pp. 1194–1198.

(*44*) Yushi Futamura et al. “Morphobase, an encyclopedic cell morphology database, and its use for drug target identification”. en. In: *Chem. Biol.* 19.12 (Dec. 2012), pp. 1620–1630.

(*45*) Jungseog Kang et al. “Improving drug discovery with high-content phenotypic screens by systematic selection of reporter cell lines”. en. In: *Nat. Biotechnol.* 34.1 (Jan. 2016), pp. 70–77.

(*46*) Karen L Huss, Pauline E Blonigen, and Robert M Campbell. “Development of a Transcreener kinase assay for protein kinase A and demonstration of concordance of data with a filter-binding assay format”. en. In: *J. Biomol. Screen.* 12.4 (June 2007), pp. 578–584.

(*47*) Chaoyang Ye et al. “DRUG-seq for miniaturized high-throughput transcriptome profiling in drug discovery”. en. In: *Nat. Commun.* 9.1 (Oct. 2018), p. 4307.

(*48*) Erin C Bush et al. “PLATE-Seq for genome-wide regulatory network analysis of high-throughput screens”. In: *Nat. Commun.* 8.1 (2017).

(*49*) Aravind Subramanian et al. “A Next Generation Connectivity Map: L1000 Platform and the First 1,000,000 Profiles”. en. In: *Cell* 171.6 (Nov. 2017), 1437–1452.e17.

(*50*) Justin Lamb et al. “The Connectivity Map: using gene-expression signatures to connect small molecules, genes, and disease”. en. In: *Science* 313.5795 (Sept. 2006), pp. 1929–1935.

(*51*) Michael B Elowitz et al. “Stochastic gene expression in a single cell”. en. In: *Science* 297.5584 (Aug. 2002), pp. 1183–1186.

(*52*) Cole Trapnell. “Defining cell types and states with single-cell genomics”. en. In: *Genome Res.* 25.10 (Oct. 2015), pp. 1491–1498.

(*53*) Sydney M Shaffer et al. “Rare cell variability and drug-induced reprogramming as a mode of cancer drug resistance”. en. In: *Nature* 546.7658 (June 2017), pp. 431–435.

(*54*) Sabrina L Spencer et al. “Non-genetic origins of cell-to-cell variability in TRAIL-induced apoptosis”. en. In: *Nature* 459.7245 (May 2009), pp. 428–432.

(*55*) Marlon Stoeckius et al. “Cell Hashing with barcoded antibodies enables multiplexing and doublet detection for single cell genomics”. en. In: *Genome Biol.* 19.1 (Dec. 2018), p. 224.

(*56*) Jase Gehring et al. *Highly Multiplexed Single-Cell RNA-seq for Defining Cell Population and Transcriptional Spaces*. 2018.

(*57*) Christopher S McGinnis et al. *MULTI-seq: Scalable sample multiplexing for single-cell RNA sequencing using lipid-tagged indices*. 2018.

(*58*) Dongju Shin et al. *Multiplexed single-cell RNA-seq via transient barcoding for drug screening*. 2018.

(*59*) Junyue Cao et al. “The single-cell transcriptional landscape of mammalian organogenesis”. In: *Nature* (2019).

(*60*) Matthew A McBrian et al. *Histone Acetylation Regulates Intracellular pH*. 2013.

(*61*) Sarah A Comerford et al. “Acetate dependence of tumors”. en. In: *Cell* 159.7 (Dec. 2014), pp. 1591–1602.

(*62*) D A Cusanovich et al. “Multiplex single-cell profiling of chromatin accessibility by combinatorial cellular indexing”. In: *Science* 348.6237 (2015), pp. 910–914.

(*63*) Junyue Cao et al. “Comprehensive single-cell transcriptional profiling of a multicellular organism”. en. In: *Science* 357.6352 (Aug. 2017), pp. 661–667.

(*64*) Leland McInnes and John Healy. “UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction”. In: (Feb. 2018). arXiv: `1802.03426 (stat.ML)`.

(*65*) Marco Jost et al. “Combined CRISPRi/a-Based Chemical Genetic Screens Reveal that Rigosertib Is a Microtubule-Destabilizing Agent”. en. In: *Mol. Cell* 68.1 (Oct. 2017), 210–223.e6.

(*66*) G Grosveld et al. “The chronic myelocytic cell line K562 contains a breakpoint in bcr and produces a chimeric bcr/c-abl transcript”. en. In: *Mol. Cell. Biol.* 6.2 (Feb. 1986), pp. 607–616.

(*67*) Emileigh K Greuber et al. “Role of ABL family kinases in cancer: from leukaemia to solid tumours”. en. In: *Nat. Rev. Cancer* 13.8 (Aug. 2013), pp. 559–571.

(*68*) Jordi Barretina et al. “The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity”. en. In: *Nature* 483.7391 (Mar. 2012), pp. 603–607.

(*69*) Chengkai Dai et al. “Loss of tumor suppressor NF1 activates HSF1 to promote carcinogenesis”. en. In: *J. Clin. Invest.* 122.10 (Oct. 2012), pp. 3742–3754.

(*70*) Laleh Haghverdi et al. “Batch effects in single-cell RNA-sequencing data are corrected by matching mutual nearest neighbors”. en. In: *Nat. Biotechnol.* 36.5 (June 2018), pp. 421–427.

(*71*) Cole Trapnell et al. “The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells”. en. In: *Nat. Biotechnol.* 32.4 (Apr. 2014), pp. 381–386.

(*72*) William Brazelle et al. “Histone deacetylase inhibitors downregulate checkpoint kinase 1 expression to induce cell death in non-small cell lung cancer cells”. en. In: *PLoS One* 5.12 (Dec. 2010), e14335.

(*73*) Jae-Seok Roe et al. “BET Bromodomain Inhibition Suppresses the Function of Hematopoietic Transcription Factors in Acute Myeloid Leukemia”. en. In: *Mol. Cell* 58.6 (June 2015), pp. 1028–1039.

(*74*) J E Brownell et al. “Tetrahymena histone acetyltransferase A: a homolog to yeast Gcn5p linking histone acetylation to gene activation”. en. In: *Cell* 84.6 (Mar. 1996), pp. 843–851.

(*75*) J Taunton, C A Hassig, and S L Schreiber. “A mammalian histone deacetylase related to the yeast transcriptional regulator Rpd3p”. en. In: *Science* 272.5260 (Apr. 1996), pp. 408–411.

(*76*) Siavash K Kurdistani. “Chromatin: a capacitor of acetate for integrated regulation of gene expression and cell physiology”. en. In: *Curr. Opin. Genet. Dev.* 26 (June 2014), pp. 53–58.

(*77*) Kathryn E Wellen et al. “ATP-citrate lyase links cellular metabolism to histone acetylation”. en. In: *Science* 324.5930 (May 2009), pp. 1076–1080.

(*78*) Aaron R Quinlan and Ira M Hall. “BEDTools: a flexible suite of utilities for comparing genomic features”. en. In: *Bioinformatics* 26.6 (Mar. 2010), pp. 841–842.

(*79*) Christian Ritz et al. “Dose-Response Analysis Using R”. en. In: *PLoS One* 10.12 (Dec. 2015), e0146021.

(*80*) Matthew E Ritchie et al. “limma powers differential expression analyses for RNA-sequencing and microarray studies”. en. In: *Nucleic Acids Res.* 43.7 (Apr. 2015), e47.

(*81*) Leif Väremo, Jens Nielsen, and Intawat Nookaew. "Enriching the gene set analysis of genome-wide data by incorporating directionality of gene expression and combining statistical hypotheses and methods". en. In: *Nucleic Acids Res.* 41.8 (Apr. 2013), pp. 4378–4391.

(*82*) Aravind Subramanian et al. "Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles". en. In: *Proc. Natl. Acad. Sci. U. S. A.* 102.43 (Oct. 2005), pp. 15545–15550.

(*83*) Itay Tirosh et al. "Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-seq". en. In: *Science* 352.6282 (Apr. 2016), pp. 189–196.

(*84*) Junyue Cao et al. "The single-cell transcriptional landscape of mammalian organogenesis". en. In: *Nature* 566.7745 (Feb. 2019), pp. 496–502.

(*85*) Jonathan S Packer et al. "A lineage-resolved molecular atlas of embryogenesis at single-cell resolution". en. In: *Science* 365.6459 (Sept. 2019).

(*86*) James A Briggs et al. "The dynamics of gene expression in vertebrate embryogenesis at single-cell resolution". en. In: *Science* 360.6392 (June 2018).

(*87*) Blanca Pijuan-Sala et al. "A single-cell molecular map of mouse gastrulation and early organogenesis". en. In: *Nature* 566.7745 (Feb. 2019), pp. 490–495.

(*88*) Jeffrey A Farrell et al. "Single-cell reconstruction of developmental trajectories during zebrafish embryogenesis". en. In: *Science* 360.6392 (June 2018).

(*89*) Nikos Karaiskos et al. "The embryo at single-cell transcriptome resolution". en. In: *Science* 358.6360 (Oct. 2017), pp. 194–199.

(*90*) Darren A Cusanovich et al. "The cis-regulatory dynamics of embryonic development at single-cell resolution". en. In: *Nature* 555.7697 (Mar. 2018), pp. 538–542.

(*91*) Caleb Weinreb et al. "Lineage tracing on transcriptional landscapes links state to fate during differentiation". en. In: *Science* 367.6479 (Feb. 2020).

(*92*) Xiaojie Qiu et al. "Reversed graph embedding resolves complex single-cell trajectories". en. In: *Nat. Methods* 14.10 (Oct. 2017), pp. 979–982.

(*93*) Patrik L Ståhl et al. "Visualization and analysis of gene expression in tissue sections by spatial transcriptomics". en. In: *Science* 353.6294 (July 2016), pp. 78–82.

(*94*) Samuel G Rodriques et al. "Slide-seq: A scalable technology for measuring genome-wide expression at high spatial resolution". en. In: *Science* 363.6434 (Mar. 2019), pp. 1463–1467.

(*95*) Kok Hao Chen et al. "RNA imaging. Spatially resolved, highly multiplexed RNA profiling in single cells". en. In: *Science* 348.6233 (Apr. 2015), aaa6090.

(*96*) Sheel Shah et al. "In Situ Transcription Profiling of Single Cells Reveals Spatial Organization of Cells in the Mouse Hippocampus". en. In: *Neuron* 92.2 (Oct. 2016), pp. 342–357.

(*97*) Je Hyuk Lee et al. “Highly multiplexed subcellular RNA sequencing in situ”. en. In: *Science* 343.6177 (Mar. 2014), pp. 1360–1363.

(*98*) Sanjay R Srivatsan et al. “Massively multiplex chemical transcriptomics at single-cell resolution”. en. In: *Science* 367.6473 (Jan. 2020), pp. 45–51.

(*99*) Mary C Regier et al. “Spatial presentation of biological molecules to cells by localized diffusive transfer”. en. In: *Lab Chip* 19.12 (June 2019), pp. 2114–2126.

(*100*) Samuel L Wolock, Romain Lopez, and Allon M Klein. “Scrublet: Computational Identification of Cell Doublets in Single-Cell Transcriptomic Data”. en. In: *Cell Syst* 8.4 (Apr. 2019), 281–291.e9.

(*101*) Laleh Haghverdi et al. “Batch effects in single-cell RNA-sequencing data are corrected by matching mutual nearest neighbors”. en. In: *Nat. Biotechnol.* 36.5 (June 2018), pp. 421–427.

(*102*) Tim Stuart et al. “Comprehensive Integration of Single-Cell Data”. en. In: *Cell* 177.7 (June 2019), 1888–1902.e21.

(*103*) Gioele La Manno et al. “Molecular architecture of the developing mouse brain”. en. July 2020.

(*104*) Hannah A Pliner, Jay Shendure, and Cole Trapnell. “Supervised classification enables rapid annotation of cell atlases”. en. In: *Nat. Methods* 16.10 (Oct. 2019), pp. 983–986.

(*105*) Tobias Pietzsch et al. “BigDataViewer: visualization and processing for large image data sets”. en. In: *Nat. Methods* 12.6 (June 2015), pp. 481–483.

(*106*) Johannes Schindelin et al. “Fiji: an open-source platform for biological-image analysis”. en. In: *Nat. Methods* 9.7 (June 2012), pp. 676–682.

(*107*) P Thévenaz, U E Ruttimann, and M Unser. “A pyramid approach to subpixel registration based on intensity”. en. In: *IEEE Trans. Image Process.* 7.1 (1998), pp. 27–41.

(*108*) Ruben Dries et al. *Giotto, a toolbox for integrative analysis and visualization of spatial expression data*.

(*109*) Graciana Diez-Roux et al. “A high-resolution anatomical atlas of the transcriptome in the mouse embryo”. en. In: *PLoS Biol.* 9.1 (Jan. 2011), e1000582.

(*110*) David Fawkner-Corbett et al. “Spatiotemporal analysis of human intestinal development at single-cell resolution”. en. In: *Cell* (Jan. 2021).

(*111*) Andreas Sagner and James Briscoe. *Establishing neuronal diversity in the spinal cord: a time and a place*. 2019.

(*112*) V Dupé and A Lumsden. “Hindbrain patterning involves graded responses to retinoic acid signalling”. en. In: *Development* 128.12 (June 2001), pp. 2199–2208.

(*113*) Suzan Abu-Abed et al. “Differential expression of the retinoic acid-metabolizing enzymes CYP26A1 and CYP26B1 during murine organogenesis”. en. In: *Mech. Dev.* 110.1-2 (Jan. 2002), pp. 173–177.

(*114*) Kirsten M Spoorendonk et al. “Retinoic acid and Cyp26b1 are critical regulators of osteogenesis in the axial skeleton”. en. In: *Development* 135.22 (Nov. 2008), pp. 3765–3774.

(*115*) Vladimir Yu Kiselev, Andrew Yiu, and Martin Hemberg. “scmap: projection of single-cell RNA-seq data across data sets”. en. In: *Nat. Methods* 15.5 (May 2018), pp. 359–362.

(*116*) B Nadarajah et al. “Two modes of radial migration in early development of the cerebral cortex”. en. In: *Nat. Neurosci.* 4.2 (Feb. 2001), pp. 143–150.

(*117*) O Marın and J L Rubenstein. “A long, remarkable journey: tangential migration in the telencephalon”. en. In: *Nat. Rev. Neurosci.* 2.11 (Nov. 2001), pp. 780–790.

(*118*) Seong-Seng Tan et al. “Cellular dispersion patterns and phenotypes in the developing mouse superior colliculus”. en. In: *Dev. Biol.* 241.1 (Jan. 2002), pp. 117–131.

(*119*) Yuji Watanabe, Chie Sakuma, and Hiroyuki Yaginuma. “Dispersing movement of tangential neuronal migration in superficial layers of the developing chick optic tectum”. en. In: *Dev. Biol.* 437.2 (May 2018), pp. 131–139.

(*120*) Junyue Cao et al. “Comprehensive single-cell transcriptional profiling of a multicellular organism”. en. In: *Science* 357.6352 (Aug. 2017), pp. 661–667.

(*121*) Darren A Cusanovich et al. “Multiplex single cell profiling of chromatin accessibility by combinatorial cellular indexing”. en. In: *Science* 348.6237 (May 2015), pp. 910–914.

(*122*) Matthew H Kaufman. *The Atlas of Mouse Development*. en. Academic Press, Sept. 1992.

(*123*) A E Petiet et al. *High-resolution magnetic resonance histology of the embryonic and neonatal mouse: A 4D atlas and morphologic database*. 2008.

(*124*) Julien Delile et al. “Single cell transcriptomics reveals spatial and temporal dynamics of gene expression in the developing mouse spinal cord”. en. In: *Development* 146.12 (Mar. 2019).

(*125*) Andrew W Senior et al. “Improved protein structure prediction using potentials from deep learning”. en. In: *Nature* 577.7792 (Jan. 2020), pp. 706–710.

(*126*) Konrad J Karczewski et al. “Author Correction: The mutational constraint spectrum quantified from variation in 141,456 humans”. en. In: *Nature* 590.7846 (Feb. 2021), E53.

(*127*) Clare Bycroft et al. “The UK Biobank resource with deep phenotyping and genomic data”. en. In: *Nature* 562.7726 (Oct. 2018), pp. 203–209.

(*128*) All of Us Research Program Investigators et al. “The “All of Us” Research Program”. en. In: *N. Engl. J. Med.* 381.7 (Aug. 2019), pp. 668–676.

(*129*) Tim Stuart et al. *Comprehensive Integration of Single-Cell Data*. 2019.

(*130*) Laleh Haghverdi et al. “Batch effects in single-cell RNA-sequencing data are corrected by matching mutual nearest neighbors”. en. In: *Nat. Biotechnol.* 36.5 (June 2018), pp. 421–427.

(*131*) Yuhan Hao et al. “Integrated analysis of multimodal single-cell data”. en. Oct. 2020.

(*132*) Marlon Stoeckius et al. “Simultaneous epitope and transcriptome measurement in single cells”. en. In: *Nat. Methods* 14.9 (Sept. 2017), pp. 865–868.

VITA

Sanjay R Srivatsan grew up in the suburbs of Atlanta, Georgia. There he attended Evansdale Elementary, Trickum Middle School and Parkview high school. Afterwards, he matriculated to the University of California, Berkeley where he graduated with a B.S. in Bioengineering from the School of Engineering. During the summers of 2010 and 2011, and during the fall semester of 2011, Sanjay spent his time at Emory University performing research under the mentorship of Dr.Periasamy Selvaraj. There he investigated methods for creating immune-stimulating microparticles derived from tumor cells with the goal of creating a prophylactic cancer vaccine. Before starting graduate school, Sanjay discovered his love for structural biology in the lab of Peter Kwong at the Vaccine Research Center in the National Institute for Allergy and Infectious Disease. There he worked with Tongqing Zhou to solve structures of broadly neutralizing antibodies specific to the HIV-1's gp120 protein. These structures were in turn used to inform the design new vaccine immunogens to elicit a HIV-1 specific immune response. When not in lab, Sanjay spends his time dreaming of rock climbing, listening to music, dreaming up new experiments, and making pizza.