EPFL

Thèse n° 10 379

Deciphering the nature of cell state transitions in single cells using quantitative modeling of temporal dynamics

Présentée le 13 juin 2024

Faculté des sciences de la vie
Unité du Prof. La Manno
Programme doctoral en biologie computationnelle et quantitative

pour l'obtention du grade de Docteur ès Sciences

par

Alex Russell LEDERER

Acceptée sur proposition du jury

Prof. P. Gönczy, président du jury
Prof. G. La Manno, directeur de thèse
Dr O. Bayraktar, rapporteur
Dr J. G. Camp, rapporteur
Prof. A.-F. Bitbol, rapporteuse

École polytechnique fédérale de Lausanne

2024

Acknowledgements

The work presented in this thesis would not have been possible without the scientific and personal support of several people who I would like to recognize here. I would like to start by thanking **Dr. Gioele La Manno** for accepting me into his research group and for placing an enormous amount of trust in me as one of his first doctoral students. Your creative yet analytical approach to tackling scientific problems greatly inspires me. I have appreciated our extensive scientific discussions and all the opportunities you have given me. Thank you for the continued mentorship and constructive feedback that has helped me grow as a scientist.

I am very grateful to the La Manno lab members. Once, we were just three! I am especially thankful for **Irina Khven**, who started together with me on this risky journey in a new lab. Together, we have experienced too many successes and struggles to count, but I could always count on you to listen. I look back fondly our tea breaks, boat days on the lake, and McDonalds evenings during the pandemic. I also want to thank **Christian Schneider** for his friendship and appreciation of tacocat, as well as **Hannah Schede** for always looking on the bright side of things and for introducing me to Lausanne's local hikes.

As time passed, the lab grew! I am so glad to see the team spirit that emerged from the newcomers (both past and present). It has been a privilege to work with all of you. This includes the current members: **Toni Herrera**, **Leila Alieh**, **Albert Dominguez Mantes**, **Alessandro Valente**, **Luca Fusar Bassini**, and **Ali Attar**. Thank you, **Daniil Bobrovskiy**, for keeping me company in the "velocity branch" of the lab! I am especially grateful for our unofficial lab weekend in Granada, which was a personal highlight of our time together (a special thanks to Toni for guiding us on his home turf). **Ercüment Çiçek**, thanks for being welcome company in an occasionally quiet office and for showing me firsthand the famous Turkish hospitality during our tour of Istanbul's cuisine and history at RECOMB.

I also want to share my gratitude to my collaborators and their labs, whose intellect and scientific rigor greatly enriched my work: **Dr. Fredrik Lanner**, **Dr. Giovanni D'Angelo**, **Dr. Felix Naef**, and **Dr. Luca Pinello**. A special thanks to **Sandra Petrus-Reurer** for sharing your research (and musical performances). I have appreciated your advice and determination!

I would like to acknowledge the members of my candidacy exam committee, **Dr. Carl Peterson**, **Dr. Pierre Fabre**, **Dr. Bart Deplancke**, and **Dr. Matteo del Peraro**, as well as my mentor **Dr. Pavan Ramdya**. I am also highly appreciative of my thesis jury for taking the time to review my work and facilitate meaningful scientific discussion: **Dr. Pierre Gönczy**, **Dr. Anne-Florence Bitbol**, **Dr. Omer Bayraktar**, and **Dr. J. Gray Camp**. I am honored and privileged to receive your valuable insights on my research endeavors.

I have been lucky to meet many talented students at EPFL who helped me to establish traditions in our young EDCB doctoral school. Thanks to the program directors: Matteo for inspiring the idea of an annual symposium and **Dr. Patrick Barth** for supporting me in making it happen – with help from **Corrine Lebet**. A special thanks to my amazing peers who made the event a reality (two times!): **Alexandre**, **Anastasia, Daphne, Jenny, Luca, Maria, Rohan, Sasha, Sibo, Sim, Simon,** and **Vamsi**. I am grateful for your friendship and hard work.

Outside of the lab, I had the opportunity to foster a passion for teaching at several courses. A heartfelt *obrigado* to the lovely **people in Braga** for offering a memorable teaching experience, and for inviting me back for an encore the following year!

My PhD times would not have been the same without the friends who started this journey with me. **Silja** and **Umair**, our board game nights, raclette dinners, and not-coffee breaks were often the best part of my week! Thank you, **Rita** and **Aspasia**, for also joining for memorable evenings and happy hours across town.

From my life before I moved to Switzerland, there are many people I want to thank. This includes my scientific mentors **Dr. Karen Arndt** and **Dr. Lars Steinmetz**, as well as **Dr. Mitch Ellison** and **Dr. Sibylle Vonesch**, for guiding me towards pursuing a PhD and sparking my scientific curiosity. I am also grateful for the other two letters of my start codon: **Logan** and **Ellie**. Thanks to my university **friends from Pittsburgh** who have kept in touch despite the geographic distance! I am appreciative of my **Long Island friends** of more than 15 years: you are the first people I rush to see when I am back in the US. I loved our virtual meetups, which we somehow coordinated across up to three different time zones. We are all so grown up!

Most of all, I want to share my love and appreciation for my family, especially **my parents** and **brother Jason**. Thank you for always supporting me at every step in my pursuit of knowledge and happiness, even when that meant the distance between us would be large and our time spent together short. I am grateful for all your visits to Switzerland, and the visits of extended family and friends. Thank you to **my grandparents**, who always reminded me that knowledge can never be taken from you, and who would be so proud to see me reach this milestone. A *vielen lieben dank* to **Philipp** and **his family** – my German family – for being my home away from home. Philipp, I am forever grateful for your continuous love and support. You make me a better person! Thank you for cheering me up whenever (and as often as) needed, for the many long walks, for the big adventures and simple evenings, and for keeping me grounded throughout all of the challenges. To everybody else who has shown me kindness or friendship that I have failed to recognize here: thank you! Exciting times lie ahead!

Lausanne, January 25th, 2024 *A.R.L.*

Abstract

Cells are the smallest operational units of living systems. Through synthesis of various biomolecules and exchange of signals with the environment, cells tightly regulate their composition to realize a specific functional state. The transformation of a cell by internal and external stimuli that alter its biomolecular composition is conceptualized as a *cell state transition* and plays a critical role in dynamic biological processes, including differentiation, development, and proliferation. Recent advances in technologies that can scrutinize cell states at a single-cell resolution, particularly single-cell RNA sequencing (scRNA-seq), offer the opportunity to assess how underlying molecular properties influence the conversion between states. However, the design of suitable computational methods to aid with interpretation of these data is an active and incomplete area of research. Here, I decode the intricate properties of cell state transitions through quantitative analyses and modeling, tackling three distinct research questions that explore the path, pace, and rules of temporal dynamics in single cells.

First, I examine cell state transitions at the population level, asking which transitions occur in a differentiation protocol where a homogeneous pool of progenitor cells is directed towards a mature cell type. I explore this question in the practical setting of an embryonic stem cell differentiation protocol to generate retinal pigmented epithelium (RPE) for treating age-related macular degeneration. Using scRNA-seq, I conclude that our protocol, rather than progressing along a linear route from stem cells to RPE, can be better explained by a divergence-convergence model of differentiation that largely recapitulates development.

Second, I investigate the pace at which cell state transitions occur, asking how the rate of the cell cycle varies across different tissues and environmental contexts and whether it can be inferred by the gene expression of an ensemble of cells. To this end, I reformulate the RNA velocity algorithm, which extrapolates future cell states from scRNA-seq data, into a unified framework with gene manifold estimation, implementing a Bayesian model for velocity inference of periodic processes. I observe variations in cell cycle speed among diverse samples and in response to chemical or genetic perturbations. I also propose an inferential framework for statistical significance testing and discover that cell cycle velocities can be approximated in real time and validated experimentally.

Third, I consider the maintenance of a steady-state biological system, asking whether the rules that govern transition probabilities among cell states can be defined using non-transcriptional modalities. To explore this, I formulate a Markov model and infer a cell transition matrix using maximum likelihood estimation from reconstructed cell lineage information in a setting where endpoint states are known but past cell states are latent. I apply the method to

characterize lipid-state switches in dermal human fibroblasts, finding a remarkable stability of states, termed lipotypes, across cell generations.

In summary, this work advances our understanding of cell state transitions for retinal progenitor differentiation, cell cycle modulations, and fibroblast plasticity, introducing new modeling strategies to tackle these dynamics with modern single-cell omics techniques.

Keywords: Cell state transitions, single-cell transcriptomics, human embryonic stem cells, retina development, age-related macular degeneration, RNA velocity, cell cycle, manifold, variational inference, Markov model

Résumé

Les cellules sont les plus petites unités opérationnelles des systèmes vivants. Grâce à la synthèse de diverses biomolécules et à l'échange de signaux avec l'environnement, les cellules régulent étroitement leur composition pour réaliser un état fonctionnel spécifique. La transformation d'une cellule par des stimuli internes et externes qui modifient sa composition biomoléculaire est conceptualisée comme *une transition d'état cellulaire* et joue un rôle essentiel dans les processus biologiques dynamiques, notamment la différenciation, le développement et la prolifération. Les progrès récents dans les technologies capables d'examiner les états cellulaires à une résolution unicellulaire, en particulier le séquençage de l'ARN de cellule unique (scRNA-seq), offrent la possibilité d'évaluer comment les propriétés moléculaires influencent la conversion entre les états. Cependant, la conception de méthodes bio-informatiques appropriées pour faciliter l'interprétation de ces données constitue un domaine de recherche actif et incomplet. Ici, je décode les propriétés complexes des transitions d'état cellulaire grâce à des analyses quantitatives et à la modélisation, en abordant trois questions de recherche distinctes qui explorent le chemin, la vitesse et les règles de la dynamique temporelle dans les cellules individuelles.

Tout d'abord, j'examine les transitions d'état cellulaire au niveau de la population, en demandant quelles transitions se produisent dans un protocole de différenciation où un pool homogène de cellules progénitrices est dirigé vers un type de cellule mature. J'explore cette question dans le cadre pratique d'un protocole de différenciation de cellules souches embryonnaires pour générer de l'épithélium pigmenté rétinien (EPR) pour traiter la dégénérescence maculaire liée à l'âge. En utilisant scRNA-seq, je conclus que notre protocole, plutôt que de progresser le long d'une voie linéaire, peut être mieux expliqué par un modèle de différenciation divergence-convergence qui récapitule en grande partie le développement.

Deuxièmement, j'étudie la vitesse à laquelle les transitions d'état cellulaire se produisent, en demandant comment la vitesse du cycle cellulaire varie selon les différents tissus et contextes environnementaux et si elle peut être déduite de l'expression génique d'un ensemble de cellules. À cette fin, je reformule l'algorithme de *RNA velocity*, qui extrapole les futurs états cellulaires à partir des données scRNA-seq, dans un cadre unifié avec estimation de la variété de gènes, mettant en œuvre un modèle Bayésien pour l'inférence de vitesse des processus périodiques. J'observe des variations de la vitesse du cycle cellulaire parmi divers échantillons et en réponse à des perturbations chimiques ou génétiques. Je propose également un cadre d'inférence pour les tests de signification statistique et découvre que les

vitesses du cycle cellulaire peuvent être approchées en temps réel et validées expérimentalement.

Troisièmement, je considère le maintien d'un système biologique à l'état d'équilibre, et investigue si les règles qui déterminent les probabilités de transition entre les états cellulaires peuvent être définies à l'aide de modalités non transcriptionnelles. Pour explorer cela, je formule un modèle de Markov et déduis une matrice de transition cellulaire en utilisant l'estimation du *Maximum Likelihood* à partir des informations de lignée cellulaire reconstruites dans un contexte où les états terminaux sont connus mais les états cellulaires passés sont latents. J'applique la méthode pour caractériser les changements d'état lipidique dans les fibroblastes dermiques humains, découvrant une stabilité remarquable des états, appelés lipotypes, à travers les générations de cellules.

En résumé, ce travail fait progresser notre compréhension des transitions d'état cellulaire pour la différenciation des progéniteurs rétiniens, les modulations du cycle cellulaire et la plasticité des fibroblastes, en introduisant de nouvelles stratégies de modélisation pour aborder ces dynamiques avec des techniques modernes d'omique des cellules uniques.

Mots clés: Transitions d'état cellulaire, transcriptomique unicellulaire, cellules souches embryonnaires humaines, développement de la rétine, dégénérescence maculaire liée à l'âge, RNA velocity, cycle cellulaire, collecteur, inférence variationnelle, modèle de Markov

List of Figures

List of Tables

List of Abbreviations

2D	two-dimensional
3D	three-dimensional
AMD	age-related macular degeneration
ANR	anterior neural ridge
ATAC	assay for transposase-accessible chromatin with sequencing
BAM	binary alignment and map
BSA	bovine serum albumin
BM	Bruch's membrane
CC	choriocapillaris
CCA	canonical correlation analysis
cDNA	complementary DNA
CELLMA	Cell-state transition Estimation by Lineage Leaf-state Markov Analysis
Cer	ceramide
ChTxB	Cholera toxin B
CRISPR	Clustered Regularly Interspaced Short Palindromic Repeats
CS	Carnegie Stage
CrNeCr	cranial neural crest
dd	differential delay
dHF	dermal human fibroblasts
DNA	Deoxyribonucleic Acid
DMEM	Dulbecco's Modified Eagle Medium
Dx	day x (i.e., 7, 14, 30, 38, 45, 60, 63, 70)
E1C3	Novo Nordisk clinical grade hESC cell line
EB	embryoid body
EMT	epithelial-to-mesenchymal transition
Endo	endothelial
Ex	mouse embryonic day x (i.e., 10, 14, 15)
FACS	fluorescence-activated cell sorting
FB	forebrain
FUCCI	fluorescence ubiquitination-based cell-cycle indication
G0	quiescent, non-cycling cell cycle phase reached through cell cycle exit
G1	gap phase of the cell cycle after M and before S
G2	gap phase of the cell cycle after S and before M
GMP	good manufacturing practice
GM1	monosialotetrahexosylganglioside
Gb3	trihexosylceramide
Gb4	globosides
GO	gene ontology
HB	hindbrain
HC	horizontal cell
HS980	human stem cell research-grade cell line
hESC	human embryonic stem cell
Hex-Cer	hexosylceramide

hrLN	human recombinant laminin
hPSC	human pluripotent stem cell
HybISS	hybridization-based in situ sequencing
InnEar	inner ear
KARO1	Karolinska Institutet clinical grade hESC cell line
KL	Kullback-Leibler
KNN	k-nearest neighbor/neighborhood
LatFold	lateral fold
LatNeEp	lateral neural epithelium
LensPlac	lens placode
LRMN	low rank multivariate normal
MALDI-MSI	matrix-assisted laser desorption/ionization mass spectrometry imaging
M	mitosis phase
MCMC	Markov chain Monte Carlo
MB	midbrain
MesCh	mesenchyme
ML	machine learning
MS	mass spectrometry
mRNA	messenger RNA
m/z	mass-to-charge
NR	neural retina
NT	non-targeted
NuMA	nuclear mitotic apparatus protein
ODE	ordinary differential equation
OS	optic stalk
OcSurEct	ocular surface ectoderm
PBS	phosphate buffered saline
PC	photoreceptor cell
PCA	principal component analysis
PEDF	pigment epithelium-derived factor
Pluri	pluripotent
PrePlac	pre-placodal epithelium
R	Pearson's correlation
R^2	squared correlation
RG	radial glia
RGC	retinal ganglion cell
RNA	ribonucleic acid
RNA-seq	ribonucleic acid sequencing
RPE	retinal pigmented epithelium
RT-qPCR	reverse transcription polymerase chain reaction
RetProg	retinal progenitor
sc	single-cell
sn	single-nucleus, single-nuclei
SD, STD	standard deviation
SD-OCT	spectral domain optical coherence tomography

SEM	standard error of the mean
S	synthesis phase
ShTxB1a	Shiga toxin 1a
ShTxB2e	Shiga toxin 2e
SM	sphingomyelins
SVI	stochastic variational inference
TF	transcription factor
TEER	transepithelial resistance
TGFβ	transforming growth factor beta
tSNE	t-distributed stochastic neighbor embedding
UMAP	uniform manifold approximation and projection
UMI	unique molecular identifier
α	alpha, transcription rate
β	beta, splicing rate
γ	gamma, degradation rate
Δν	delta-nu, batch effort for gene expression coefficients
ν	nu, gene expression Fourier series coefficients
νω	nu-omega, angular speed Fourier series coefficients
ω	omega, angular speed, velocity
ϕ	phi, cell cycle phase

Table of Contents

1. Introduction

1.1. Cell states and cell state transitions

1.1.1. Reductionism in biology and the *Zellenstaat* concept

The tendency to seek out analogies that can distill profound observations into approachable terms has long been a core practice of scientific inquiry. A classic analogy is the comparison of a cell to a factory: the organelles in a cell are similar to the operating units in a factory, such as the nucleus (control center), ribosomes (production machinery), and mitochondria (power plants). This reductionist view aims to break down systems into their underlying pieces in order to better describe how they work.

A reductionist's mindset has dominated biological research for decades (Van Regenmortel, 2004). In the twentieth century, this point was conveyed by Francis Crick, recipient of the Nobel Prize for his breakthrough work on the structure of DNA, when he claimed in his lecture "Of molecules and men" that *"the ultimate aim of the modern movement in biology is to explain all biology in terms of physics and chemistry"* (Crick, 1966). At the beginning of the twenty-first century, shortly after the first drafts of the sequenced human genome were completed, researchers again argued that in order to understand biology, one needed to *"learn how to speak the language of the genome fluently"* (Hub Zwart, 2007). This analogy suggests that knowing the meaning of the words and sentences (or genes and functional regions) of the genome is essential to describing intricate biological phenomena.

One strategy to interrogate the role of an individual gene is the "loss of function" or "knockout" approach, in which a gene is silenced or deleted, and the effects on the molecular or physical composition of cells are studied. Gene knockdowns have been systematically applied to characterize gene lethality (Giaever & Nislow, 2014; Ross-Macdonald et al., 1999) as well as the role of genes in patterning the developing embryo (Mocellin & Provenzano, 2004; Nasevicius & Ekker, 2000; Nüsslein-Volhard & Wieschaus, 1980; Zimmer et al., 2019). Novel methodological techniques, particularly the CRISPR/Cas9 system (Bock et al., 2022; Jinek et al., 2012), have accelerated perturbation screening efforts by making it easier to target a particular gene or functional region for removal. The impact of a gene knockdown is usually evaluated by measuring changes to the molecular status of a cell and the concentrations of specific biomolecules.

Another way to elucidate the links between genes (genotype) and cellular features (phenotype) is to monitor gene activity over time and evaluate how the status of a cell changes in relation to temporal differences in its molecular contents. This can help to group similar cells by their shared characteristics, and to figure out which groups of cells are collectively needed

to carry out all functional requirements of an organism. Modern biological research involves describing the mechanisms that define how cells operate, and mapping cells to the larger system in which they belong.

German physician Rudolf Virchow considered yet another analogy, this time between cells and human beings, when he coined the concept of a ***cell state***, or "*Zellenstaat*." As the father of modern pathology and cell theory, Virchow compared cells of a body to citizens of a society, with each entity playing an important role to support the "economy of the organism" (Maehle, 2011; Mulas et al., 2021). In his published *Cellularpathologie* theory from the 1850s, Virchow wrote:

"The character and the unity of life cannot be found in one particular single point of higher organization, such as the human brain, but only in particular, constantly recurring arrangements, which every single element owns. From this follows that the composition of a larger body, the so-called individual, always results in a kind of social arrangement, [and] represents an organism of a social kind, where a mass of single existences is dependent on each other, but in such a manner that each element […] has a particular activity for itself, and that each, although it may receive the stimulus for its activity from other parts, still is itself the origin of its actual work…" (Virchow, 1858).

At the time, little was understood about the composition of a cell, so Virchow could not have known the full aptness of his analogy. In the following years, discoveries found even smaller components (i.e., organelles) and molecules inside cells that carry out "a particular activity" to coordinate essential biological processes (Burbridge & Adrain, 2022). This again transformed the concept of a "*Zellenstaat"* to refer to a cell's detailed molecular configuration (Ferry, 2019).

Before knowing that nucleic acids were the chemical language of genes, scientists used non-specific stainings of nuclear material or other histological methodologies to discriminate between dissimilar cell states (Morris, 2019). Later, with a clearer understanding of the genetic code and the flow of genetic information during transcription and translation, the concept of a cell state evolved to be defined by the abundance or concentration of specific molecular units: DNA, RNA, and proteins. Experimental measurement of these quantities was first made possible by classical molecular biology techniques such as Western and Northern blotting, and later with next-generation sequencing technologies. Research using these methods helped to annotate cell states using a vast number of molecular parameters, including gene expression, chromatin accessibility, histone modifications, and protein levels.

With a better grasp of the role of individual genes and the entities that regulate them, more complex molecular features are now applied to describe cell types that would appear indistinguishable from other states based on physical or morphological attributes alone (Reynolds, 2007).

1.1.2. Cell state transitions and the progenitor cell state

Crucially, an underlying motif throughout advancements in the interpretation of cell state has been the recognition that cells change over time. Given that complex organisms arise from a single cell, yet contain many distinct cell states with varying molecular phenotypes, it is irrefutable that cells need to transform and transition between states. Cells change to accomplish temporal processes such as differentiation, activation, and proliferation, and these conversions are triggered in response to internal (i.e., genetic modifications, gene regulation) and external (i.e., environmental stress, cell-to-cell communication, signaling cues) factors. The impact of transitions on a cell's molecular configuration highlights the dynamic nature of living systems.

Initial studies of ***cell state transitions*** led to speculation of a precursor or progenitor cell capable of transforming into multiple specialized cell types. Sketches from physician Artur Pappenheim illustrated the concept of cell state transitions during blood hematopoiesis (**Fig. 1.1**), in which a "*Mutterzelle*," or mother cell, differentiates into diverse cell types of the blood (Maehle, 2011). This progenitor would eventually become known as the pluripotent stem cell. Similar discoveries were made in developmental biology: the embryologist Caspar Friedrich Wolff observed that the early embryo comprised of three germ layers and argued, contrary to the consensus at the time, that an embryo of differentiated cells arose from undifferentiated cells (Kiecker et al., 2016). This later inspired scientists such as Hans Spemann to manipulate the germ layers by physically moving cells to different regions of the embryo; he found that transplanted cells induced other non-transplanted cells to change their state, but only at specific developmental stages (Spemann & Mangold, 1923). Walter Vogt sought to continue Spemann's *"conquest of an uncharted territory of knowledge"* by tracing cell lineages with dyes during their differentiation (Gilbert, 2007; Hsu, 2015).

The accumulation of a body of research showing that stem cells are capable and responsible for generating all of the complex cell types of a mature organism fueled many important considerations about cell state transitions, some which continue to be debated (Mulas et al., 2021). For example, the question of the permeance and reversibility of a cell state remains unanswered. Some research posits that during a linear process such as differentiation, cells transit through various commitment points, or intermediate states, at

which cells become irreversibly committed to a particular fate. However, specific chemical factors and external cues can induce the reversion of fully differentiated cells back to an earlier progenitor state (Shi et al., 2017; Takahashi & Yamanaka, 2006). Naturally, this spurs another question of whether the ***route*** taken by cells to transition between two cell states even matters, or if different paths can be followed to reach the same end point (Brackston et al., 2018). The timing and ***rate*** of cell state transitions is also key to inducing the correct cell fate, as shown by Spemann and many others (De Robertis, 2006; Garcia-Ojalvo & Bulut-Karslioglu, 2023; Grove & Monuki, 2020; Pera & Rossant, 2021).

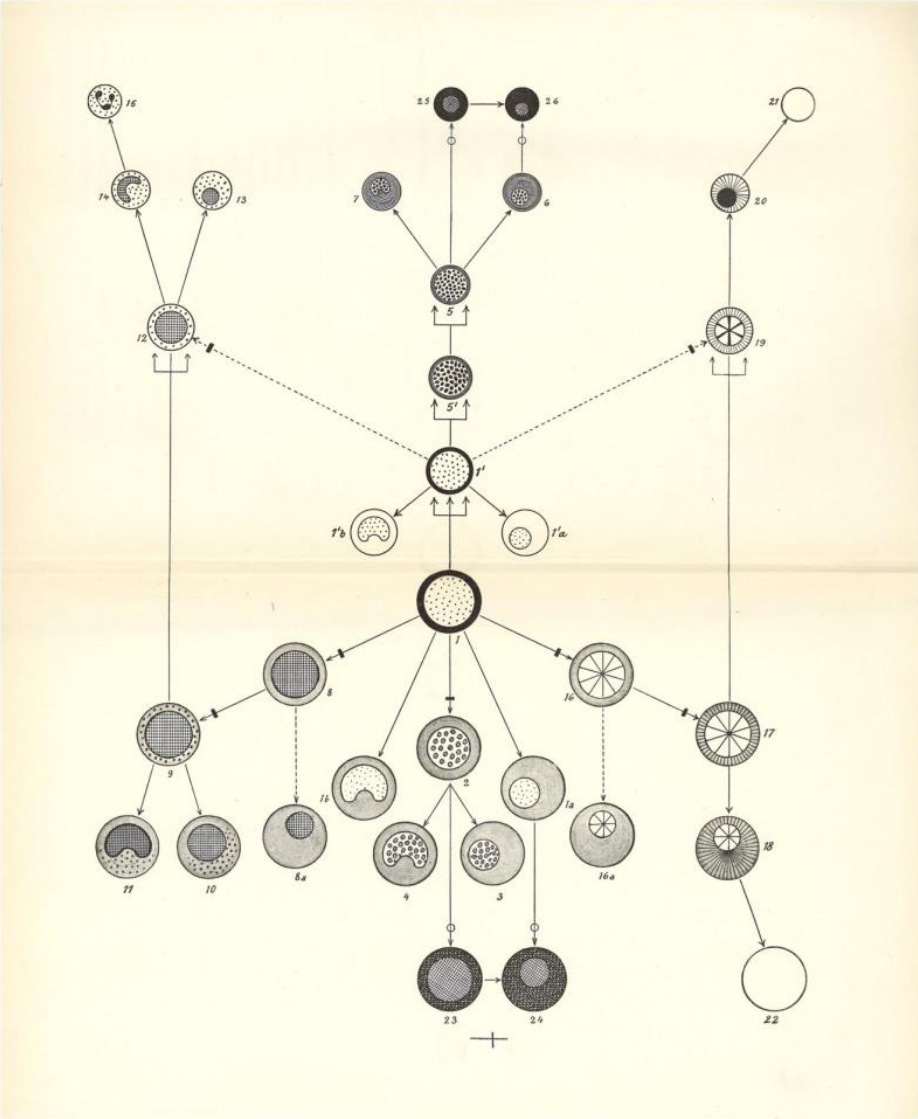


Figure 1.1. Pappenheim's early drawing of stem cell differentiation. He proposed a "Mutterzelle" (black circle; center) as the source of four different branches of mature cell types, which are produced only after transiting through numerous intermediate cell states. This image was taken from: Maehle: "Ambiguous Cells: The Emergence of the Stem Cell Concept in the Nineteenth and Twentieth Centuries." Notes Rec R Soc Lond (2011). The original sketch by Pappenheim was published as: Pappenheim: "Atlas der menschlichen Blutzellen." Gustav Fischer in Jena (1905).

Furthermore, there is recognition that cell states fall along a continuous spectrum defined by multiple axes of variation. For example, many functionally distinct cell types undergo phases of rapid proliferation, which requires fine tuning of expression for a specific ensemble of genes that is shared across all proliferative cell types, no matter how different in spatial location or functional context. Thus, cell states can be multifaceted and are often defined by co-occurring but independently operating genetic programs (Miroshnikova et al., 2023).

1.2. Modeling cell state transitions at the level of biological systems

1.2.1. Waddington's phase space and epigenetic landscape

One of the challenges when studying transitions taken by cells based on changes to their molecular composition is representing the effects of those numerous components with a straightforward and interpretable model. C.H. Waddington contemplated this problem on the scale of embryos in development and proposed a conceptual model that recognized popular physics methods introduced by his contemporaries. In this analogy, the complex dynamics of embryonic development are simplified to a trajectory within a high-dimensional "phase space," with axes corresponding to the relevant features (**Fig. 1.2**). Waddington proposed those features could be the genes or their "chemical tendencies" (Waddington, 1957). However, the full extent to which genes directly encode proteins was, at that time, not quite understood. Contrary to those in the field of physics, Waddington does not consider the rate of change as axes of his phase space; furthermore, the realization that the ***concentrations*** of gene products control the progression of an embryo through the feature space is missing from his original analogy.

On this phase space, molecular changes occur within specific components of the embryo according to certain rates, which define those component's trajectory in time. Waddington describes his phase space as follows:

"In the study of development, we are interested not only in the final state to which the system arrives, but also in the course by which it gets there…we must fall back on a mode of expression which may be called geometrical rather than algebraic. A system containing many components can be represented by a point in multidimensional space, the co-ordinates of the point in each dimension representing the measure of a particular component. A space of this kind is known as a phase space. As the composition of the system changes the point will move along a certain trajectory" (Waddington, 1957).

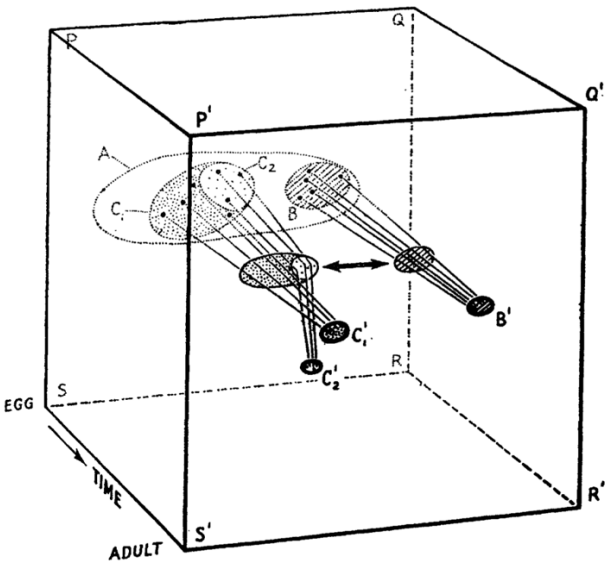


Figure 1.2. Waddington's phase-space diagram of development. With this sketch, Waddington proposed that changes to the individual components (i.e., C_1, C_2) of the early embryo (or egg) occur during development in a high-dimensional space (PQRS). Each coordinate of that space maps to a particular component of the embryo, and those components change their coordinates along a certain trajectory in response to gene expression changes until reaching an adult. The end points of each component (i.e., C'_1, C'_2) will differ in the final high-dimensional space (P'Q'R'S'). In recent years, scientists have repurposed this model to consider cells as the individual components and states as their positions in high-dimensional space. This diagram is taken from *"The strategy of genes"* by C.H. Waddington (1957).

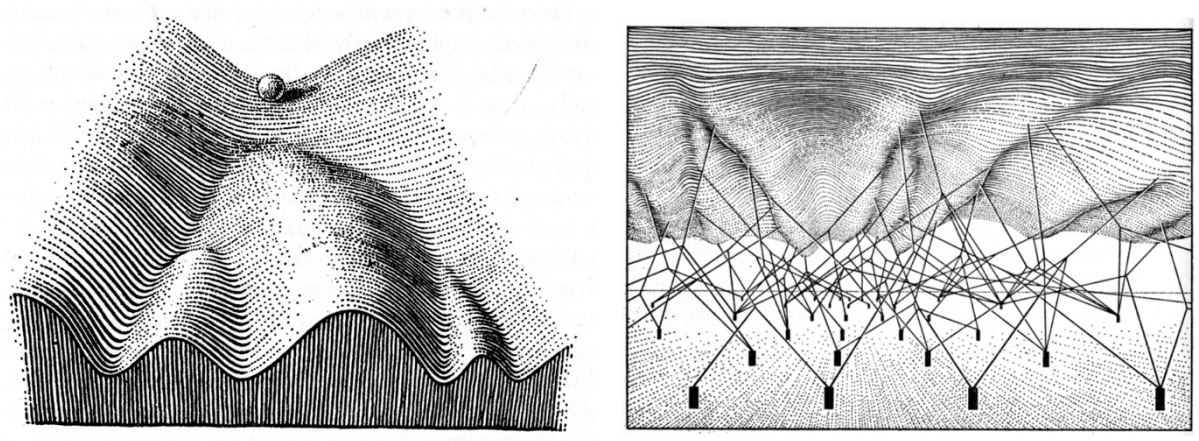


Figure 1.3. Waddington's epigenetic landscape. Left: Waddington's sketch entitled *"Part of an Epigenetic Landscape"* in which an embryo, or ball, rolls down the developmental landscape defined by hills and valleys. Right: Waddington's sketch entitled *"The complex system of interactions underlying the epigenetic landscape"* in which genes, or strings, exert pull on the landscape, changing its topography and influencing the path taken by the ball towards the bottom. This figure was taken from *"The strategy of genes"* by C.H. Waddington (1957).

Scientists further extended this analogy by considering cells as the components of the embryo, states as the positions in phase space, and cell state transitions as the trajectories in phase space over time. Waddington himself sought to simplify his proposal of a high-dimensional phase space by suggesting a low-dimensional representation called an "epigenetic landscape". In this metaphor, a developing embryo is likened to a ball rolling along a path in a landscape defined by numerous hills and valleys. Under the influence of genes, the embryo is guided through various intermediates until reaching a particular destination, or developed individual (**Fig. 1.3**). Genes and their expressed corresponding molecules (not fully understood by Waddington) pull on the landscape from underneath the surface, affecting the path taken by the embryo from a unicellular to multicellular state.

In recent years, this metaphor has been reworked to liken the ball to a stem cell, rather than an entire embryo. In this case, the contour of the hills separates different possible cell state transitions and guides a progenitor cell towards a terminal cell type (**Fig. 1.4, top**).

Although Waddington imagined embryos traversing this landscape in a one-directional manner, as in development, one can think of a wider set of valid, yet less canonical, cell state transitions in various biological contexts (**Fig 1.4, bottom**). This includes cell reprogramming (**Chapter 2**) or the interconversion between semi-stable states (**Chapter 4**). Likewise, the forces affecting the path taken through the landscape do not necessarily need to be dictated by genes alone (as in an autonomous system), but can also be viewed as changes in response to environmental stimuli or external signaling. This means that biological systems, which are heavily influenced by factors outside their internal environment, are often non-autonomous systems.

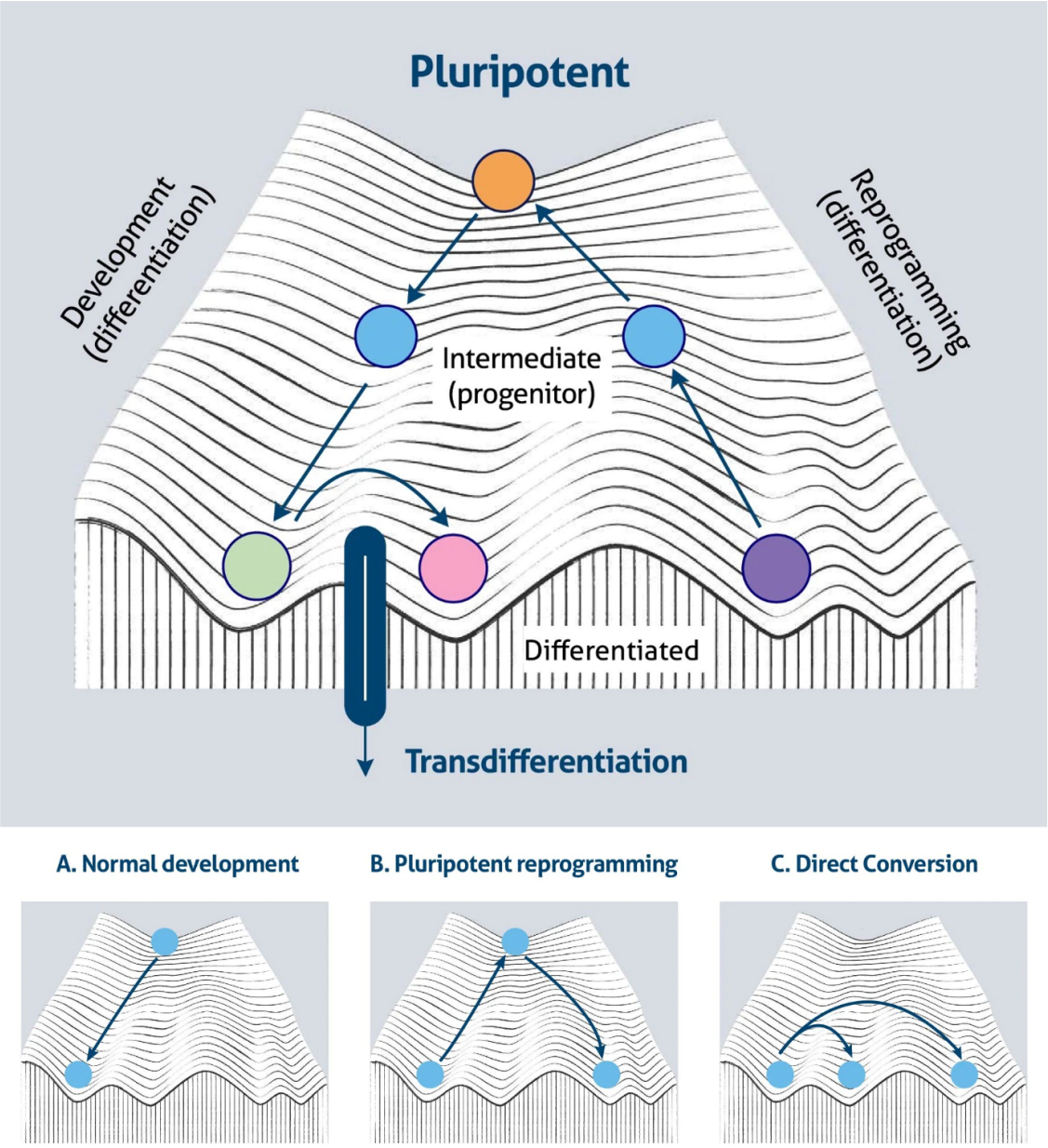


Figure 1.4. A modern interpretation of Waddington's epigenetic landscape. Top: adjusted schematic of Waddington's landscape in which an undifferentiated stem cell (orange) traverses through intermediate cell states (blue) before reaching a mature, differentiated cell state (green, pink, or purple). Bottom: Waddington's landscape was conceptualized as one-directional, as in development, but can also be extended to account for pluripotent stem cell reprogramming as well as the interconversion among differentiated, semi-stable cell states. This figure was adapted from *"Cell fate commitment and the Waddington landscape model"* by Proteintech, available at the following link as of 24.01.2024: https://www.ptglab.com/news/blog/cell-fate-commitment-and-the-waddington-landscape-model/.

1.2.2. Formulating cell state transitions with ordinary differentiation equations

To mathematically formulate changes in the concentration of "important materials" that define an embryo's transition along a developmental trajectory in phase space, Waddington proposed the use of an autonomous set of ordinary differential equations (ODEs). ODEs describe the rate of change for a variable, such as the expression of a particular gene, and how that rate is linked to the variable itself, as well as others within the system. In the autonomous system case, the potential effect on the key variables by an external perturbation or control is assumed to be negligible. An early application of ODEs to study gene regulation was for the inducible lac-operon system discovered by François Jacob and Jacques Monod in *E. coli*, for which they received the Nobel Prize in 1965 (Jacob & Monod, 1961).

In the lac-operon system, a cell changes between two enzyme-catalyzing states depending on the nutrients present. When glucose, the preferred energy source, is present in the environment, the operon is repressed. However, when only lactose is available and not glucose, the operon is induced and an adjacent set of three structural genes is expressed to make three enzymes: beta-galactosidase, beta-galactoside permease, and beta-galactoside transacetylase. Beta-galactosidase hydrolyzes lactose to galactose and glucose, thereby providing energy to the bacteria even in a glucose-deficient environment (Jacob & Monod, 1961; Lewis, 2011).

Following the identification of this gene regulatory module, Monod sought to model the enzyme synthesis rate of beta-galactosidase using differential equations (Jacob & Monod, 1961; Monod et al., 1952). In this situation, the enzyme synthesis rate was controlled by an external inducer molecule, lactose. These efforts inspired scientists to model numerous other types of inducible and repressible systems in bacteria (Santillán & Mackey, 2008). For example, J.S. Griffith proposed that two cell states could exist stably, as attractors, under certain circumstances based on a gene regulated by a positive feedback loop (Griffith, 1968).

Extrapolating these ideas to model complex systems involving many genes, one could describe the progression of a cell through numerous cell states over time as an autonomous dynamical system, expressed as a system of ODEs describing the rate of change of one gene in relation to the other genes. The relationships (transitions) between a cell's potential attractor states are represented by the joint behavior of these equations. However, most cell states are not described by a fully autonomous systems because they are influenced by fluxes in the environment and changes in nearby cells (Ferrell, 2012; Jaeger & Monk, 2014; Rand et al., 2021). Despite this stochasticity, dynamical systems models using ODEs can be tweaked and applied to infer how a cell's state evolves over time in response to both changes in gene expression and other external stimuli. With the right data, these classes of models can provide

valuable insights into the paths of differentiation and conditions under which cells transition between states (Sáez et al., 2022).

1.3. Harnessing single-cell transcriptomics to profile cell states

Shifting away from models like Waddington's, which were mostly conceptual, towards models like Monod's, which were viewed as a means for interpreting data, requires methods that can measure molecular composition and abundances in cells. In recent years, next-generation sequencing technologies, which enabled the high-throughput collection of molecular measurements in cells, provided a breakthrough in the complexity and ambition of modeling efforts in biology.

1.3.1. Single-cell RNA-sequencing and omics atlases

High-throughput sequencing technologies have revolutionized how biologists study cell state transitions, particularly in dynamic settings such as stem cell differentiation, embryonic development, and disease (Metzker, 2010). The earliest approaches to profile gene expression at the whole-transcriptome level, such as microarrays (Niemitz, 2007; Schena et al., 1995) and RNA-sequencing (Lister et al., 2008; Mortazavi et al., 2008; Nagalakshmi et al., 2008), enabled global phenotyping of bulk cell populations for the first time. With these technological advancements came the rise of modern systems biology, with increased computational modeling and analysis of gene expression patterns, transcriptional circuits, and regulatory networks in the context of cell state transitions (Alon et al., 1999; Bergmann et al., 2003; Blanpain & Simons, 2013; Eisen et al., 1998; Klein & Simons, 2011; Shen-Orr et al., 2002). One example is the Zeisel equations, which, inspired by Monod's rate equations for enzymatic activity, model RNA expression dynamics in response to an acute stimulation, or in temporally recurrent situations such as during circadian cycles (Zeisel et al., 2011).

Unfortunately, microarrays and RNA sequencing obtain average gene expression measurements over millions of cells; unless there is a cell sorting step, these bulk techniques mask an underlying heterogeneity of the sample (Kulkarni et al., 2019). Therefore, the invention of single-cell RNA-sequencing (scRNA-seq) about 15 years ago (F. Tang et al., 2009) has facilitated a transformative leap in cell state characterization by enabling the study of gene expression in individual cells, exposing an intricate heterogeneity within tissues (Hashimshony et al., 2012; Islam et al., 2011; Klein et al., 2015; Macosko et al., 2015). Commercialized protocols involve isolating and suspending cells in separate microfluidic droplets, each containing a unique cellular barcode. The cell's mRNA is then reverse-transcribed into cDNA, and a unique molecular identifier (UMI) is incorporated to measure

expression in absolute terms. Transcripts are then amplified in a pooled setting, which generates a library for sequencing. Finally, computational preprocessing of the sequenced reads maps them to individual cells and genes, allowing a detailed evaluation of cell-to-cell variation (Haque et al., 2017; Jovic et al., 2022).

Rapid technological advancements in scRNA-seq have greatly increased the number of cells that can be profiled at a time (**Fig. 1.5**). It is now common to measure the transcriptome of hundreds of thousands of cells in just one study (Svensson et al., 2020). Recent experimental techniques can simultaneously recover information about other cell attributes at single-cell resolution, including chromatin accessibility (Buenrostro et al., 2015; Cao et al., 2018; Song Chen et al., 2019; S. Ma et al., 2020; Zhu et al., 2019), DNA methylation (Ahn et al., 2021; Karemaker & Vermeulen, 2018), histone modifications (Bartosovic et al., 2021; Kaya-Okur et al., 2019), and surface protein levels (Labib & Kelley, 2020; Petrosius & Schoof, 2023; Stoeckius et al., 2017).

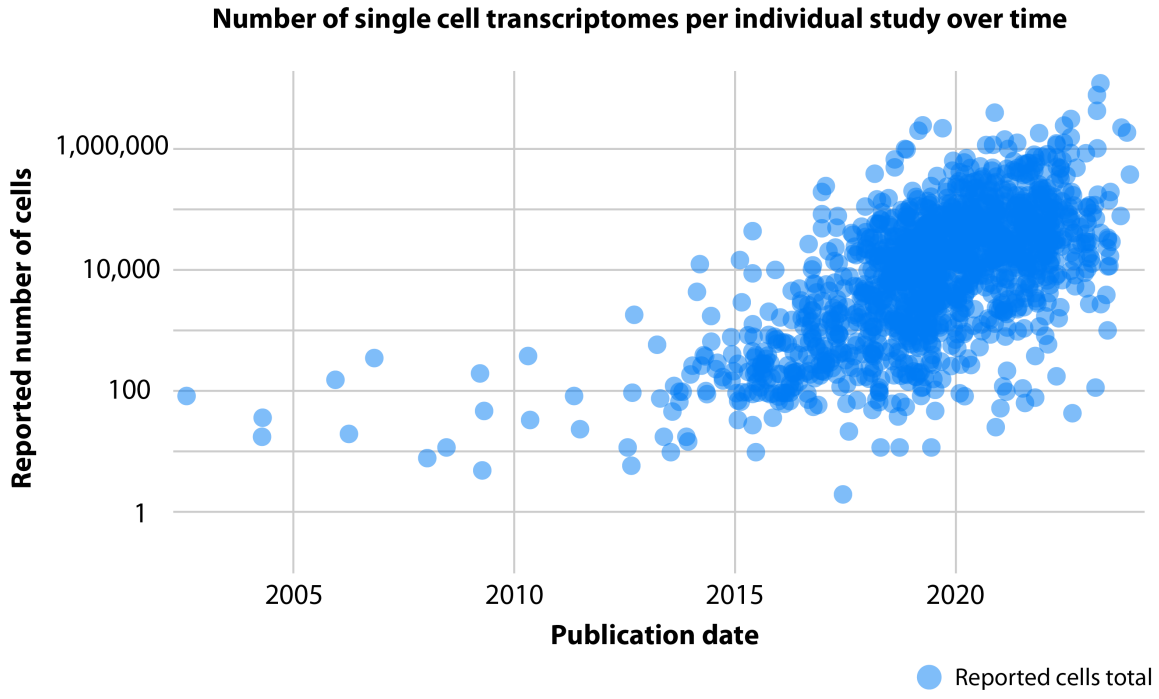


Figure 1.5. Scatter plot of the number of single-cell transcriptomes obtained by research studies over the past 20 years. With improvements in experimental and computational methodologies, the number of single cells or nuclei that can be profiled using scRNA-seq has grown at an exponential rate. Each dot (n=1,927) represents one study that generated new single-cell transcriptomic data. The y-axis uses a logarithmic scale. This figure is adapted from the "Single Cell Studies Database" as of January 2024 and was originally published in the following article: *Svensson, da Veiga Beltrame, Pachter. "A curated database reveals trends in single-cell transcriptomics". Database (2020).*

Many research efforts have sought to comprehensively define and annotate all of the cell types in various tissues and species at single-cell resolution. The large datasets generated by these projects are referred to as "single-cell atlases" and are critical resources for the scientific community (Lindeboom et al., 2021; Stephen R. Quake, 2022; Rood et al., 2022). Considering the analogy that compares cells in a tissue to humans in a society, from which Virchow's "*Zellenstaat*" originates, these atlases are the equivalent of a census for a particular biological system, thoroughly recording the properties of unique entities making up a larger organ or organism.

Single-cell atlases have aided scientists with testing hypotheses about changes in cell states during cancer (Siyuan Chen et al., 2023), neurodegenerative disease (Piwecka et al., 2023; Pozojevic & Spielmann, 2023), and immune system disorders (Ginhoux et al., 2022). Here, diseased or aberrant tissue can be surveyed by scRNA-seq and its cell states compared to those of a healthy reference. Single-cell transcriptomics is also regularly applied to identify intermediate cell states that occur at specific time points in embryonic development and drive the tissue diversification (Haniffa et al., 2021; Vinsland & Linnarsson, 2022). Examining these transient cell states can foster key insights into the trajectories along which progenitor cells progress into mature cell types (**Chapter 1.4**). Taken together, single-cell omics technologies provide an exceptional opportunity to discern cell types across both mature adult tissues and in dynamic biological systems. Nonetheless, some technical and computational complexities continue to challenge data analysis, particularly in settings with transitioning cells; this is an active area of method development by the single-cell genomics community (Lähnemann et al., 2020).

1.3.2. Addressing data sparsity at single-cell resolution

A first obstacle when using scRNA-seq data to monitor cell state transitions is the abundance of observed zeros. This inherent sparsity means that for many genes in each cell, no UMIs are captured. In part, these zeros in the data are caused by technical noise that is acquired during molecule barcoding, library preparation, and sequencing. To address this, one can improve the sample quality, change the single-cell platform, or increase the sequencing depth. However, the absence of expression for a gene may allude to certain biological significance that cannot be overlooked: in the case of evolving cell states, the absence of expression could be due to the gradual downregulation of that particular gene.

Some methods try to take advantage of the binary attributes of scRNA-seq (and other omics) data to identify highly variable features, cluster cell types, and perform differential expression analysis (Bouland et al., 2021; P. Qiu, 2020). Nonetheless, most standard

approaches instead seek to transform count data in a way that reduces its sparsity, such as with k-nearest neighbor (KNN) smoothing to average UMIs among nearby cells in Euclidean space (Luecken & Theis, 2019). For populations of steady state cell types, KNN imputation can be highly effective at reducing noise; in a sample containing closely-related intermediate cell states, the procedure can risk blending together information and obscuring meaningful expression differences. In fact, a recent study that systematically compared data imputation methods found that these approaches did little to improve performance during downstream analyses, compared to no imputation at all (Hou et al., 2020).

1.3.3. Dimensionality reduction and gene expression manifolds

A second obstacle with scRNA-seq data is its multi-dimensionality. The feature space of a single-cell dataset is tens of thousands of genes, and there has been a recent exponential increase in the observation space size (**Fig. 1.5**). Importantly, while single-cell experiments measure a high-dimensional space, most biological activity unfolds within a significantly smaller subspace (Linderman, 2021). One reason for this is evolution, which has altered genes so that they operate together as part of coordinated expression and regulatory programs (Pope & Medzhitov, 2018). Thus, the number of independent axes representing the majority of the variance in single-cell data is much lower than even the subspace.

Dimensionality reduction techniques capture the most significant components of this subspace and describe dynamic biological processes in a low-dimensional space. The most popular strategy to achieve this is principal component analysis (Wold et al., 1987). A few dozen components are often enough to capture all relevant space on which biology unfolds; this means that most phenomena are restricted to a more compact space within the entire high-dimensional gene expression space. The concept of such a manifold, which describes a space that is locally flat, yet globally can fold and curve in a complex manner, is highly relevant to characterization of cellular trajectories, such as differentiation, that are embedded within high-dimensional scRNA-seq data (Gunawan et al., 2023; Moon et al., 2018).

Furthermore, cell transitions occur along a path constrained to the real gene expression space (Sáez et al., 2022). This enables certain properties of transitions to be modeled, including the rate of change in gene expression. Over time, this rate should describe a cell's progression on a trajectory that is tangent to the gene expression manifold, ensuring that all transited cell states lie within the feasible biological space. However, the way to formulate a model to describe such velocities as consistent to a low-dimensional subspace is non-trivial. Ultimately, considering the true feature space of single-cell data offers a geometric perspective to the study of the coordination of cell states.

Dimensionality reduction techniques can be best harnessed to model smooth and continuous cell state transition processes, but there are some situations in biology with clear attractor (stable) cell states. In these scenarios, one might more pragmatically model the manifold using discrete models. For example, Markov models describe transitions between states as occurring according to defined probabilities that are memoryless, or independent of previous transitions. To learn the properties of cell state transitions in high-dimensional single-cell data of systems at homeostasis, one could design Markov models in which the edges between states are latent variables (Chu et al., 2017; P. B. Gupta et al., 2011).

1.3.4. Studying cell state transitions with static snapshots

A third obstacle to more dynamic modeling of single cell data is that these technologies provide only a snapshot of gene expression at the time of sampling. In other words, the same cell cannot be easily measured more than once. As a consequence of this limitation, the first wave of studies to use single-cell genomics data focused on the identification and classification of complex cell types in heterogeneous tissues across species. These early single-cell studies lack a deep temporal characterization of the obtained data and focus on characterizing discrete cell types, with limited analyses centered on mapping relationships between cell types and other more transient states.

Recent technological developments such as Live-seq (W. Chen et al., 2022) show it is possible to sample the transcriptome of the same individual cell at multiple time points, but these approaches are difficult to implement, low-throughput, and restricted to *in vitro* settings. Hence, experimental or computational techniques designed to extract information on collective cellular dynamics from snapshots provide a more promising avenue for inferring temporal dynamics (Ding et al., 2022). Efforts to infer the continuous dynamics of gene expression, rather than merely provide descriptive insights, characterize an emerging second phase in the single-cell transcriptomics field.

1.4. Emergence of single-cell temporal-omics approaches

In this section, I introduce the rise of single-cell computational methods that enable examination of gene expression dynamics along cell state transitions. Some of the text here is adapted from a review article I previously wrote on this topic (Lederer & La Manno, 2020). I highlight tools that allow for estimation of the future state of a cell, particularly pseudotime trajectory inference and RNA velocity. These methods establish an important framework with which to perform the genome-wide study of change in individual cells and have greatly influenced the research landscape of single-cell transcriptomics in recent years (Lähnemann

et al., 2020; Morris, 2019; Yu & Scolnick, 2022). Taken together, the dynamic modeling of single-cell data has given rise to a second wave of studies incorporating *temporal-omics* approaches, focused on inferring changes to biological systems with respect to time (**Fig. 1.6**).

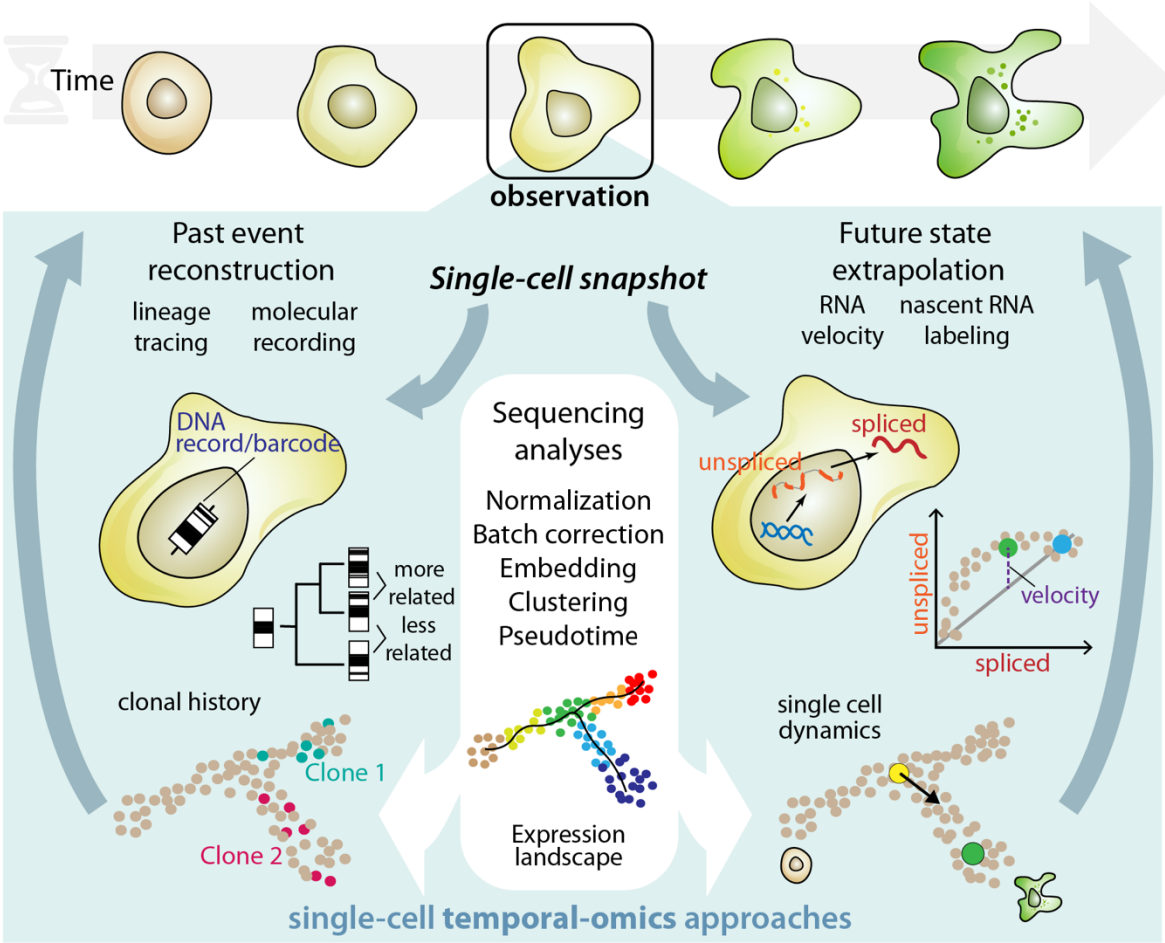


Figure 1.6. Schematic representation of *single-cell temporal-omics* approaches. Given the snapshot obtained from single-cell RNA sequencing, data analyses enable characterization and classification of the gene expression landscape in a heterogeneous population of cells. Recent methods allow for the extrapolation of future gene expression states (right) or reconstruction of past cellular events (left). Together, these approaches permit greater inference of the temporal changes within a single cell while relying on measurement at a single time point. In this thesis, I will focus on the application of future state extrapolation methods, particularly RNA velocity, to the study of cell state transitions. This figure and caption are adapted from the following article: *Lederer & La Manno. "The emergence and promise of single-cell temporal-omics approaches". Current Opinion in Biotechnology (2020).*

1.4.1. Pseudotime trajectory inference

Sampling a tissue with scRNA-seq generates a snapshot of cells in gene expression space along a biological process, revealing regions occupied by nascent, intermediate, and mature cell types. The first computational methods to represent cells as spanning expression

space and harness this information to infer dynamic cell state transitions were Monocle (Trapnell et al., 2014) and Wanderlust (Bendall et al., 2014). These tools assume that each cell is at a slightly different moment of the biological process of interest; unlike live cell approaches, they require variation among cell types within the measured population and do not infer cellular trajectories using change in a single cell. By exploiting the asynchronicity of cells within tissue samples from a single time point, cells can be ordered according to an internal time that summarizes how the expression profile of the average cell changes. If the major component of variability is affiliated with progression of development or disease, these algorithms effectively summarize that axis of variation and cluster temporally related cells into a single-cell state. This is essentially the concept of pseudotime, a variable estimated from the distribution of gene expression data and used as a proxy for cellular progression through a biological phenomenon (Deconinck et al., 2021). Since these pioneering early methods, a large number of trajectory inference algorithms were subsequently developed (Saelens et al., 2019) using mathematical concepts such as principal graphs (Albergante et al., 2018; Wolf et al., 2019), minimum spanning trees (Trapnell et al., 2014), nearest neighborhood graphs (Baran et al., 2019; Grün, 2020), or diffusion maps (Haghverdi et al., 2016) to summarize variation and aggregate temporally-related cells.

Pseudotime trajectory inference methods have grown in sophistication and accuracy, but there are three common limitations. First, these tools assume that the major component of variability between expression profiles is time, yet variation in gene expression can be caused by multiple factors, including spatial patterning, histological organization and morphogenetic gradients (Cheng et al., 2019; Joost et al., 2016). In scenarios where these other mechanisms mix with temporal changes, some traditional trajectory approaches may overfit. Moreover, technical artifacts due to noise between technologies and batches can also create components of variability that are wholly non-biological in nature (Luecken et al., 2022; Ranek et al., 2022; Tran et al., 2020).

Second, for trajectory inference methods to work well, experiments should obtain a balanced sample from the distribution of cell states traversed over time. However, a bias towards more stable or slowly-changing states may cause transient states to be distorted or missed (Tritschler et al., 2019). This limitation can be counteracted with improved experimental designs, but this is challenging when studying a poorly defined or highly complex process such as the pathology of disease or embryonic development. It is also non-trivial to determine how often to sample a developmental process, particularly if tissue is difficult to acquire. How balanced a dataset is may also partially depend on the speed of the process being modeled.

Third, while in scenarios where the fitted pseudotime does not depart from the latent time of the process, its relation with gene expression can offer valuable biological insight, these models can only be interpreted as describing statistical expectation rather than the real path taken by cells. Individual cells may have moved back and forth, around the mean, in gene expression space, and pseudotime analysis would not be able to reveal it (Huang, 2012; Weinreb et al., 2018). Trajectory inference is therefore limited because the time component of longer-range trajectories is only inferred from population averages. More recent methods such as RNA velocity, metabolic labeling, and molecular recording begin to address these limitations by estimating dynamics from information obtained from the single-cell measurement (Battich et al., 2020; Erhard et al., 2019; Gorin et al., 2023; La Manno et al., 2018; Lear & Shipman, 2023).

1.4.2. RNA velocity in single cells

Modeling the rate of a biological process using the causal relationship between two cellular entities was first proposed in 1952 by biochemist Jacques Monod in the context of enzyme-catalyzing reactions (**Chapter 1.2**). These models were revisited in the microarray era with the Zeisel equations for the purpose of studying RNA expression dynamics in response to an acute stimulation, or in temporally recurrent situations such as during circadian cycles (Zeisel et al., 2011). Since the amount of cellular unspliced mRNA determines the future quantity of spliced mRNA, these equations can describe the rate of change in mRNA accumulation.

This concept was applied to scRNA-seq data with RNA velocity, which distinguishes between immature (unspliced) and mature (spliced) molecules using intronic reads that are removed prior to translation (La Manno et al., 2018). RNA velocity is formulated from a first-order system of differential equations describing gene-specific splicing and degradation rates and is represented as a vector field in low dimensional space. The original velocity approach, implemented in *velocyto*, assumes that after transcriptional upregulation, gene expression reaches a steady-state level prior to downregulation. RNA velocities are calculated by finding a linear fit for the steady-state spliced-unspliced ratio and measuring the residuals between that obtained fit and measured expression levels. Steady-state RNA velocity assumes a shared splicing rate across all genes and requires that a steady-state cell population is sampled in the experiment (**Fig. 1.7**).

Unfortunately, these assumptions can be violated: there might not be a steady-state mature cell population present in a dataset, and the splicing rate can vary widely between different genes. To address these limitations, an improved *scvelo* framework (Bergen et al.,

2020) reformulates RNA velocity using a likelihood-based dynamical model, allowing generalization of RNA velocity to biological systems that do not contain cells at steady state. The dynamical model estimates the velocity parameters (transcription, splicing, and degradation rates) and a cellular latent time iteratively using expectation-maximization.

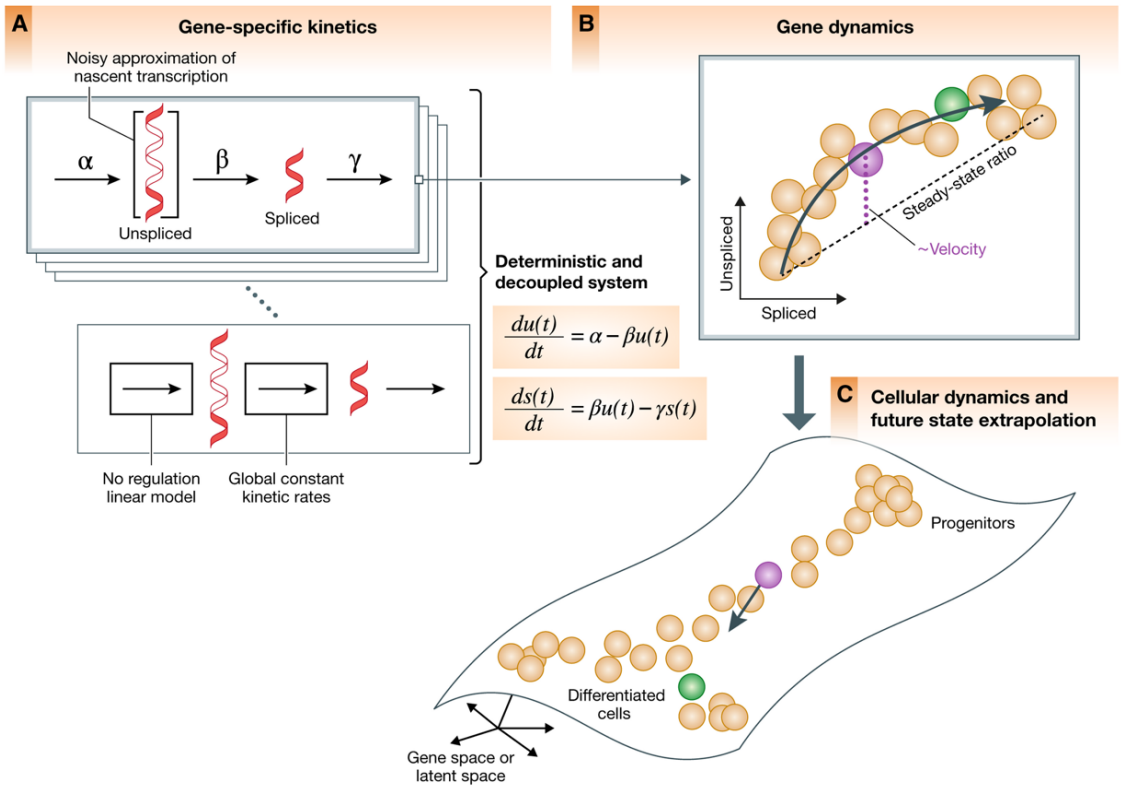


Figure 1.7. Overview of the steady-state model of RNA velocity estimation. (A) The metabolic life cycle of an RNA molecule can be described in three stages: (1) the transcription of an unspliced pre-mRNA molecule from DNA, (2) the removal of introns from unspliced RNA to generate a mature spliced RNA molecule, and (3) the eventual degradation of the spliced molecule. This process can be modeled using a set of differentiation equations and gene-specific kinetic variables for transcription (α), splicing (β), and degradation (γ). (B) The observed temporal delay between unspliced and spliced expression can be visualized on a gene-wise phase portrait. In the steady-state model, it is assumed that the unspliced-spliced ratio is roughly equal to the gene-wise degradation rate. RNA velocity is computed by calculating a linear fit of the steady-state ratio and measuring the residuals between the fit and a particular cell. Importantly, this step assumes a steady-state population is detected in the sample, which is not always the case. (C) In a last step, RNA velocity is projected onto a lower dimensional embedding as a vector field, allowing extrapolation of the progression towards future cell states along a biological path. This figure is taken from the following article: *Bergen, Soldatov, Kharchenko, and Theis. "RNA velocity – current challenges and future perspectives". Molecular Systems Biology (2021).*

Ongoing efforts have sought to extend or repurpose the RNA velocity equations to incorporate multiple omics modalities, including chromatin accessibility (Burdziak et al., 2023; C. Li et al., 2022; S. Ma et al., 2020; Tedesco et al., 2022), protein levels (Gorin et al., 2020), histone modifications (Bartosovic & Castelo-Branco, 2023), metabolically-labeled nascent

RNA (Erhard et al., 2019; Hendriks et al., 2019), and transcription factor regulation (J. Li et al., 2023). Other methods incorporate RNA velocity vectors into pseudotime inference (Lange et al., 2022) or infer velocities with a unified latent time (M. Gao et al., 2022), neural networks (Z. Chen et al., 2022; Cui et al., 2024; S. Li et al., 2023), differential geometry (X. Qiu et al., 2022), or representation learning (Qiao & Huang, 2021). Most recently, several works have proposed models for RNA velocity that are implemented using variational inference or a Bayesian framework (Aivazidis et al., 2023; Gayoso et al., 2023; Gu et al., 2022; Maizels et al., 2023; Qin et al., 2022). The large quantity of RNA velocity methods proposed by the single-cell community in recent years emphasizes the demand for tools to temporally evaluate single-cell data. Moreover, it has established RNA velocity as a crucial component of the scRNA-seq analysis toolkit, alongside pseudotime trajectory inference.

1.4.3. Recording clonal information and past transcriptional states

Lineage tracing describes the relatedness of individual cells formed in a tissue according to deterministic representations of cell type lineages (Kester & van Oudenaarden, 2018). Single-cell methods reconstruct the mitotic kinship of cells using inducible recording systems in which barcodes are incorporated into the genome and read out during sampling by sequencing or imaging-based approaches (Kebschull & Zador, 2018). Cells with similar barcodes are considered to have clonal ancestry. Recent technologies incorporate both the labeling of cells and scRNA-seq as encoded in mRNA expression profiles (Chan et al., 2019; Guo et al., 2019; McKenna et al., 2016; Raj et al., 2018; Spanjaard et al., 2018).

The two most common experimental approaches for lineage tracing are viral barcoding (Kong et al., 2020; Weinreb et al., 2020) and the CRISPR/Cas9 system (Alemany et al., 2018; Chow et al., 2021; Z. He et al., 2022; McKenna et al., 2016; F. Schmidt et al., 2018). For example, MEMOIR (Frieda et al., 2017) can directly record cellular events in response to the activation of signaling pathways, using sequential single-molecule FISH to read out the status directly. Alternatively, CAMERA (W. Tang & Liu, 2018) uses CRISPR/Cas9 to record the duration of signaling events in mammalian cell culture systems. Yet another method is CellTagging (Kong et al., 2020), which uses lentiviral libraries and combinatorial cell indexing to label cells without any genetic manipulations, reading out barcodes as polyadenylated mRNA in parallel to gene expression information during scRNA-seq.

Future improvements for lineage tracing methods include increasing barcode capture rates, decreasing mutation levels, and designing more comprehensive analysis tools to reconstruct lineages (Yao et al., 2022). One limitation to these approaches is that they only recover indirect knowledge of cell relationships, rather than direct lineages, which could be

obtained by live-cell microscopy (Z. He et al., 2022). Another drawback is that lineage tracing itself does not record specific information about the specific features of past cellular states driving cell state changes. To address this, transcriptome-memory tools that store information about past gene expression profiles themselves could offer more insight into past cell states beyond lineage information. Such a concept of molecular memory has been applied in bacteria with Record-seq, which captures expressed RNA sequences and converts them into DNA for storage (F. Schmidt et al., 2018, 2022). The development of similar tools for transcriptome-recording in eukaryotic cells would greatly enhance the study of past cell states, but admittedly comes with many technical challenges (Wagner & Klein, 2020).

1.5. Scope of the thesis

In this thesis, my goal is to disentangle the fundamental temporal aspects of cell state transitions through quantitative analysis and models. This objective is structured around three central research questions, each of which examines cell state transitions through a unique lens. My findings are available in some public format outside of this thesis, either as part of a published article or a preprint, and they have been adapted here accordingly. These details as well as individual author contributions are indicated in the preface for each chapter. Collectively, these works demonstrate a notable contribution towards the phenomenological characterization and computational modeling of the ***path***, ***pace***, and ***rules*** of cell state transitions (**Fig. 1.8**).

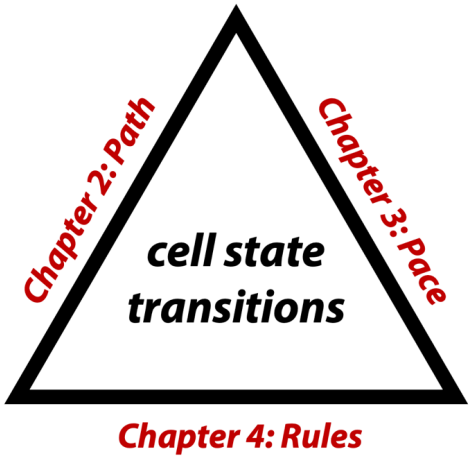


Figure 1.8. The three dimensions of cell state transitions. My thesis covers three research questions that each tackle a different aspect of cell state transitions. Chapter 2 (Path): how do cell transitions unfold during differentiation? Chapter 3 (Pace): does the rate at which cells transit through states change? Chapter 4 (Rules): what are the rules that govern transition probabilities?

1.5.1. Mapping the path taken by embryonic stem cells towards retinal tissues in targeted differentiation protocols

In Chapter 2, I start by exploring cell state transitions from a global perspective, asking which transitions are made on the route taken by human embryonic stem cells (hESC) when they differentiate in culture towards a target cell type. I examine this question in differentiation protocols that are designed to generate cells for replacement therapy against age-related macular degeneration (AMD). AMD is a major cause of vision loss and affects millions of individuals worldwide (Gehrs et al., 2006). Specifically, the "dry form" of AMD is characterized by atrophy of the ***retinal pigmented epithelium*** (RPE) cell monolayer (Hadziahmetovic & Malek, 2021). In healthy tissue, the RPE forms a boundary between densely packed photoreceptors, which perform phototransduction, and the choroid, which supplies blood to the retina (**Fig. 1.9A**). RPE plays an essential role in photoreceptor renewal via phagocytosis, nutrient and waste transport, and protection against oxidative stress (Yang et al., 2021). With an increased accumulation of extracellular debris, membrane thinning, and immune activity, photoreceptor renewal is compromised and the RPE layer breaks down (**Fig. 1.9B-C**) (Wong et al., 2022). While vision loss is usually gradual, it can vary in speed and severity among individuals, and few strategies for treatment are available (Gehrs et al., 2006; Hadziahmetovic & Malek, 2021).

Nonetheless, the directed differentiation of hESCs into neuroepithelial-derived cell types is a promising method to generate healthy tissue and arrest AMD progression (Choudhary et al., 2017; da Cruz et al., 2018; Hazim et al., 2017; Higuchi et al., 2017; Michelet et al., 2020; Alvaro Plaza Reyes et al., 2016; R. Sharma et al., 2019). Most *in vitro* protocols attempt to recapitulate the sequence of cues that trigger differentiation and bias cell fate towards RPE. Notable efforts have been made towards maximizing purity and compatibility of the produced cell pool and shortening protocol duration (A. Plaza Reyes et al., 2020). However, attention to integrity of the final product has overshadowed characterization of the intermediate stages appearing before a final mature cell state is reached. This knowledge gap has persisted due to the lack of a technology that can systematically distinguish mixed phenotypes and off-target effects from cell heterogeneity and with which to perform a quantitative cell-state comparison to relevant physiological references.

Single-cell RNA sequencing (scRNA-seq) can fill that gap, enabling molecular phenotyping of heterogeneous cell populations across all intermediate states; recent studies have revealed that *in vitro* differentiation can either traverse alternative paths through gene expression space or a similar program of intermediates as in development (Kulkarni et al., 2019; Kumar et al., 2017). Evaluation of differentiation protocols with scRNA-seq has helped

characterize the developmental trajectories of hESC-derived tissues (Cuomo et al., 2020; McCracken et al., 2020) and comprehensive cell atlases have decomposed retinal complexity at fetal and adult stages. However, a comparative understanding of the similarities between transient retinal cell states arising during development and hESC-derived intermediates is still needed (Collin et al., 2019; Cowan et al., 2020; Hu et al., 2019; Lukowski et al., 2019; Mao et al., 2019; Menon et al., 2019; Rheaume et al., 2018; Shekhar et al., 2016; Sridhar et al., 2020; Voigt et al., 2019). Describing the transient stages of differentiation protocols is crucial to understanding whether cells reach a mature cell type by mimicking the developmental processes that generate the same cell type diversity and whether any off-target cell lineages or other contaminant populations arise. Moreover, examining the plasticity of these end-state populations is essential to better design protocols to obtain the desired cell types, and to evaluate the safety and efficiency of a stem cell product for clinical use.

Here, I apply single-cell transcriptomics to characterize the path hESC take during a specific 60-day RPE differentiation protocol (A. Plaza Reyes et al., 2020). To obtain a wide-ranging assessment of cell state transitions, I analyze scRNA-seq data from six time points during a 2D monolayer differentiation performed in three different cell lines, as well as from similar time points in a 3D embryoid body protocol. Furthermore, I map the *in vitro* transcriptional states to newly acquired *in vivo* references of human fetal and adult retinas. I also probe the plasticity of a subpopulation of neuroepithelial progenitors using a neuronal differentiation scheme and following *in vivo* transplantation in the rabbit retina. With this study, I offer insight into the developmental trajectory and cell state phenomenology of 2D and 3D differentiations. I also demonstrate how modern single-cell atlases can be applied to evaluate critical stages of differentiation protocols with high accuracy and to interpret the complexity of the emerging cell states.

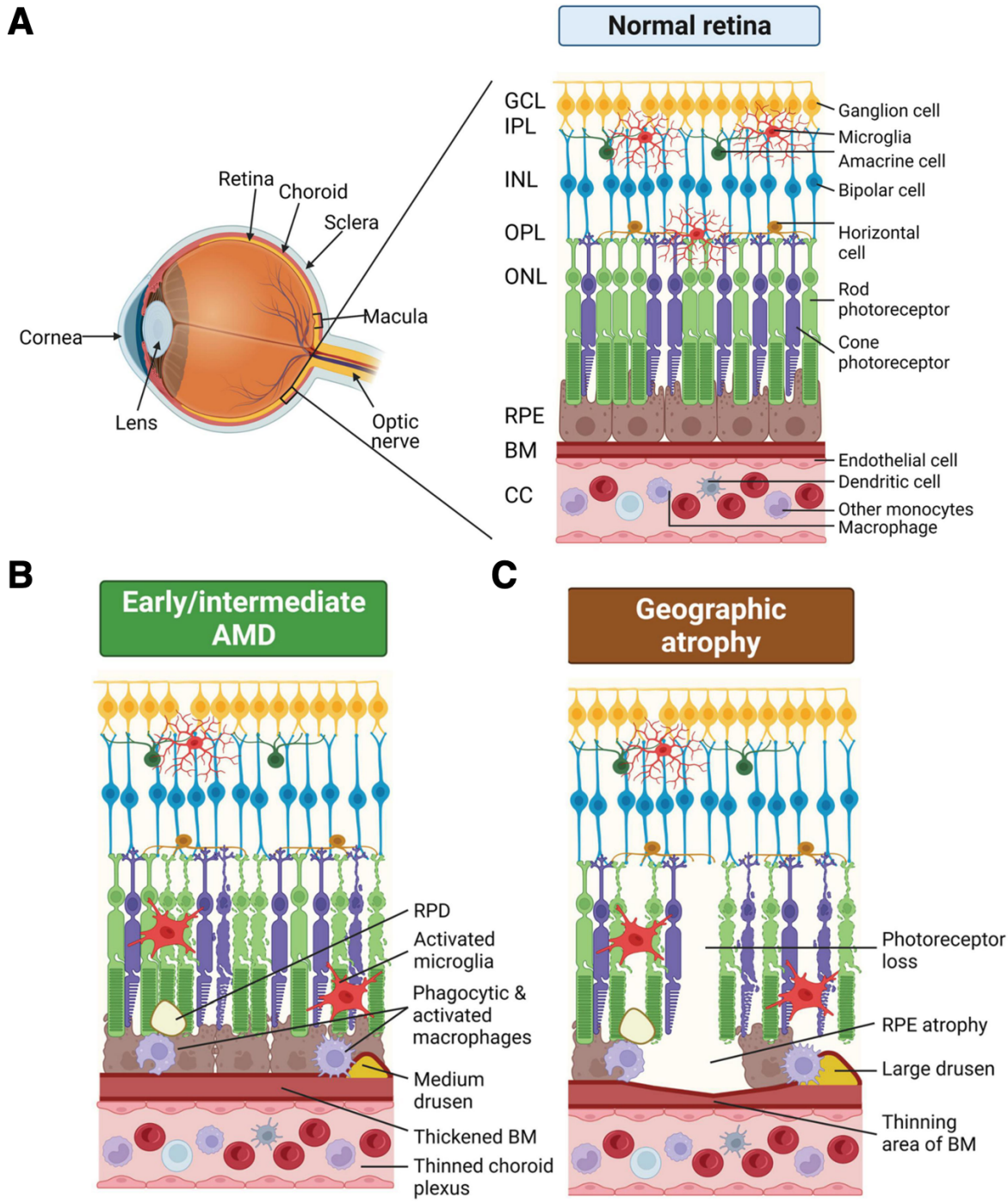


Figure 1.9. Schematic depicting the progression of age-related macular degeneration in the retina. **(A)** An overview of the anatomical structure of the healthy retina, prior to onset of AMD. Retinal pigmented epithelium (RPE) is indicated in brown; rod and cone photoreceptors are indicated in green and purple, respectively. The choroid is divided into two parts: the vascular supply, or choriocapillaris (CC), and Bruch's membrane (BM). **(B)** The early stages of AMD are characterized by thickening of the BM, accumulation of extracellular waste, and activation of immune cells in the retina. **(C)** The late stages of AMD, also known as geographic atrophy, are characterized by the dramatic loss of photoreceptors and RPE. This figure is adapted from: *Wong, Ma, Jobling, Brandli, Greferath, Fletcher, and Vessey. "Exploring the pathogenesis of age-related macular degeneration: A review of the interplay between retinal pigment epithelium dysfunction and the innate immune system". Frontiers in Neuroscience (2022).*

1.5.2. Measuring the pace at which cells transition during the cell cycle across tissues using probabilistic models

Besides the path taken by cells during differentiation, the timing itself of cell transitions is highly coordinated to ensure proper development (Mulas et al., 2021). Therefore, in Chapter 3, I consider the speed at which cells move between states along their trajectories, asking whether this can be modeled using oscillations in unspliced and spliced gene expression.

As previously noted, ***RNA velocity*** has emerged as a popular technique to reconstruct temporal information from static single-cell snapshots and estimate the rate of change in gene expression (**Chapter 1.4**). However, there are known limitations to the algorithm, including that it uses nearest-neighbor smoothing to approximate expectations on the counts and that it relies on non-linear dimensionality reduction to bring the high dimensional velocity vector onto a two-dimensional embedding (Bergen et al., 2021; Gorin et al., 2022). Another major limitation is that most RNA velocity models do not perform velocity estimation jointly on all genes. In other words, individual gene-wise velocities are aggregated into a global latent time, producing a geometrically inconsistent velocity vector, with gene-specific components on different timescales. Finally, it is difficult to establish a ground truth for cell transition rates against which to benchmark RNA velocity algorithms, prohibiting many sensitivity analyses that are typically applied to new computational tools.

To address these problems, I reformulate RNA velocity and gene expression manifold estimation in a unified framework designed to track one-dimensional periodic manifolds. This facilitates model validation and application to the cell cycle. The cell cycle is the most prominent periodic process in biology and its pace is crucial for tissue homeostasis and diseases such as cancer (Matthews et al., 2022). In healthy tissue, the cell cycle occurs in four sequential phases (G1, S, G2, and M) and is responsible for duplication of the genetic material and cell division (**Fig. 1.10**). In scRNA-seq data obtained from embryonic samples, the cell cycle is usually a major axis of variation (Chervov & Zinovyev, 2022; Satija et al., 2015). For example, radial glia cells of the brain and neuronal progenitors in the retina both proliferate at different speeds along temporal and spatial axes (Davis & Dyer, 2010; Ohnuma & Harris, 2003).

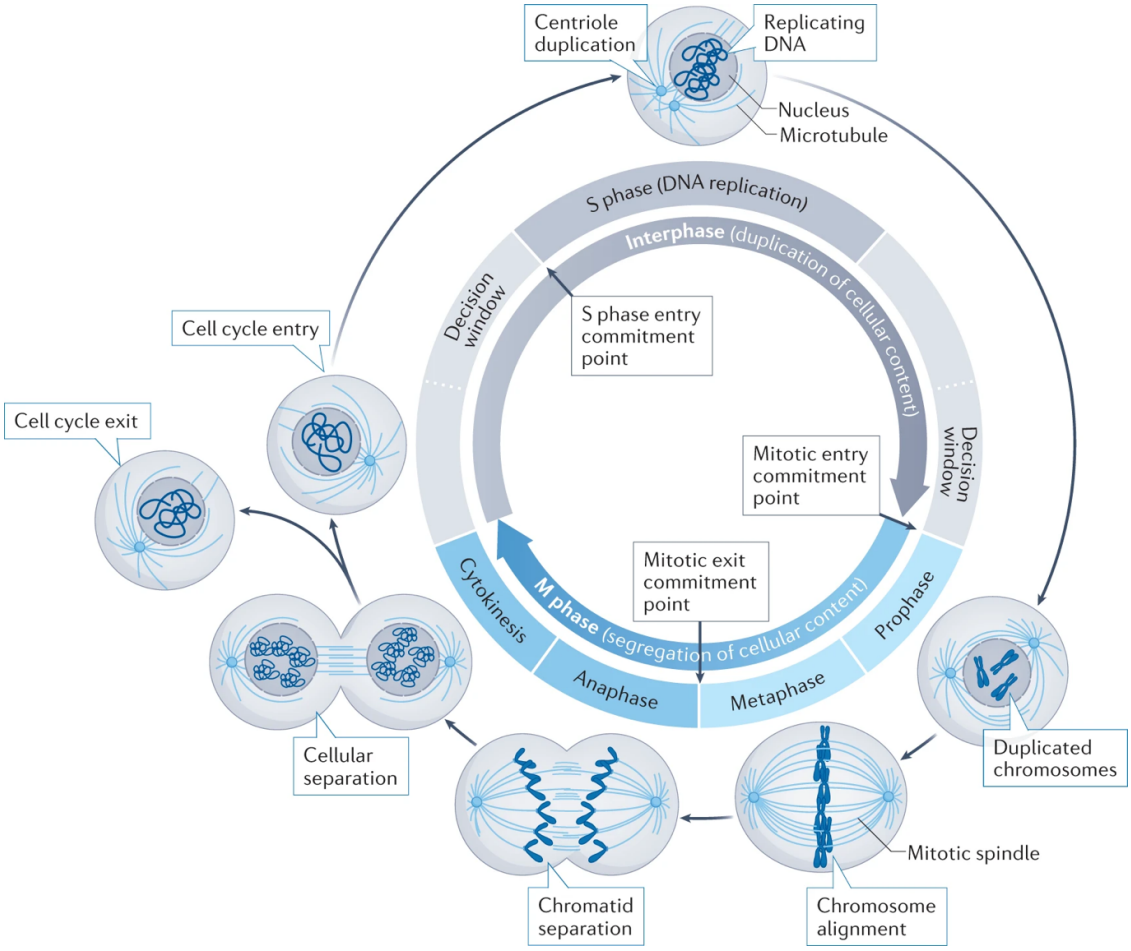


Figure 1.10. Schematic of the cell cycle. The cell cycle is divided into interphase, when cellular content is duplicated, and mitosis, when cellular content is segregated into two. Interphase can be further divided into S phase, when the DNA is duplicated, and two gap-phases (G1 and G2, labeled as decision window). The G1 gap phase prior to S phase as well as the G2 gap phase after S phase are crucial cell checkpoints and tightly regulated by numerous well-known genes. After cell division in mitosis (M), two daughter cells are generated, and it is possible for one of both of them to exit the cell cycle and enter a quiescent phase (G0). This figure is taken from: *Matthews, Bertoli, and de Bruin. "Cell cycle control in cancer." Nature Reviews Molecular Cell Biology (2022).*

My formulation of RNA velocity is implemented as a Bayesian framework named *VeloCycle*, which operates on raw counts and is solved using variational inference in Pyro. This strategy allows for sensitivity analyses and demonstrates one- and multiple-sample statistical testing. I benchmark RNA velocity inference on multiple *in vitro* and *in vivo* datasets and demonstrate its utility as a useful resource for a more reflective RNA velocity analysis.

1.5.3. Monitoring the rules that govern cell lipid-state transitions in stable cell populations of dermal fibroblasts

In most biological systems, both the path and pace of cell state transitions are pre-defined. However, there are some circumstances in which transitions recur in order to maintain

a certain proportion of cell states; modeling these systems is difficult from static snapshot data, as there is no obvious “source” or “sink” cell state. Variation in these systems may be defined not by transcriptomic features, but rather by other omics modalities that are significantly more difficult to measure at a single-cell resolution. Unfortunately, existing computational methods are poorly suited to address these types of recurring cell dynamics, which do not occur in embryonic development but rather in biological systems at homeostasis.

In Chapter 4, I consider cell state transitions necessary for maintaining a steady-state biological system, asking whether it is possible to determine the rules defining cell transitions using non-transcriptional modalities. To examine this, I consider the system of dermal human fibroblasts (dHF), which are known to fluctuate between a multitude of cell states to perform diverse cellular functions to facilitate wound healing, activate pro-inflammatory signatures, and proliferate (Adler et al., 2020; Driskell & Watt, 2015; Philippeos et al., 2018; Rognoni et al., 2018). However, the rules governing this established heterogeneity of cellular states needed in dHFs are not fully understood.

Lipid measurements at single-cell resolution can only be obtained by imaging, rather than sequencing. Consequently, I propose a new computational approach that models cell state transitions as memoryless Markov chains and estimates a transition probability matrix among lipid-defined states from coupled time-lapse microscopy and endpoint toxin-staining read outs. I find that dHFs cultured at steady state transit among dynamic sphingolipid confirmations called ***lipotypes*** that are propagated across cell generations and correspond to phenotypic states called lipotypes.

2. Molecular profiling of stem cell-derived retinal pigment epithelial cell differentiation established for clinical translation

Sandra Petrus-Reurer[a,b,c,*], Alex R. Lederer[d,*], Laura Baqué-Vidal[a,b], Iyadh Douagie, Belinda Pannagel[e], Irina Khven[d], Monica Aronsson[c], Hammurabi Bartuma[c], Magdalena Wagner[a,b], Andreas Wrona[f], Paschalis Efstathopoulos[f], Elham Jaberi[f], Hanni Willenbrock[f], Yutaka Shimizu[f], J. Carlos Villaescusa[f], Helder André[c], Erik Sundström[g], Aparna Bhaduri[h], Arnold Kriegstein[h], Anders Kvanta[c], Gioele La Manno[d,†], Fredrik Lanner[a,b,i,†]

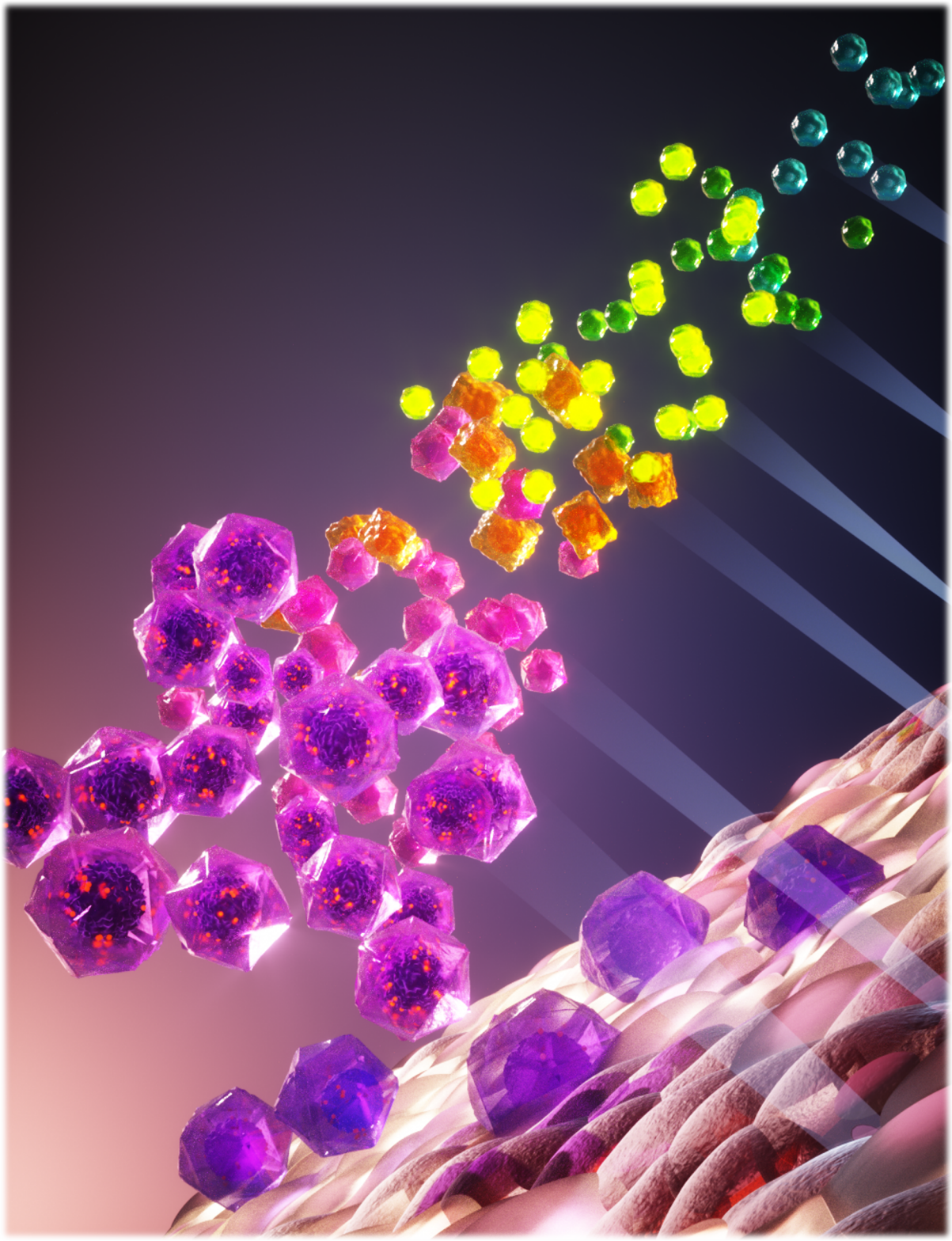
(a) Department of Clinical Sciences, Intervention and Technology, Karolinska Institutet, 17177 Stockholm, Sweden.
(b) Gynecology and Reproductive Medicine, Karolinska Universitetssjukhuset, 14186 Stockholm, Sweden.
(c) Department of Clinical Neuroscience, Division of Eye and Vision, St. Erik Eye Hospital, Karolinska Institutet, 11282 Stockholm, Sweden.
(d) Laboratory of Neurodevelopmental Systems Biology, Brain Mind Institute, School of Life Sciences, Ecole Polytechnique Fédérale de Lausanne (EPFL), 1015 Lausanne, Switzerland.
(e) Center for Hematology and Regenerative Medicine, Department of Medicine, Karolinska Institutet, 17177 Stockholm, Sweden.
(f) Cell Therapy R&D, Novo Nordisk A/S, Måløv DK-2760, Denmark.
(g) Department of Neurobiology, Care Sciences and Society, Karolinska Institutet, 17177 Stockholm, Sweden.
(h) Department of Neurology, University of California, San Francisco, CA, USA; Eli and Edythe Broad Center for Regeneration Medicine and Stem Cell Research, University of California, San Francisco, CA, USA.
(i) Ming Wai Lau Center for Reparative Medicine, Stockholm node, Karolinska Institutet, 17177 Stockholm, Sweden.

* Authors contributed equally
† Corresponding authors: Gioele La Manno (gioele.lamanno@epfl.ch) and Fredrik Lanner (fredrik.lanner@ki.se)

Published in *Stem Cell Reports* (Petrus-Reurer et al., 2022) and available at: https://doi.org/10.1016/j.stemcr.2022.05.005.

I am co-first author of this research work.
Featured on the cover of June 2022 issue.

An earlier preprint version can be found at: https://doi.org/10.1101/2021.01.31.429014.



Artwork 2.1. Cell state transitions during retinal pigmented epithelium differentiation. Human embryonic stem cells travel through transcriptomically distinct cell states on the route towards mature retinal pigmented epithelium. We hope to use these cells for transplantation therapies to halt age-related macular degeneration. This illustration was designed by Ella Maru Studio and featured as the cover *Stem Cell Reports* for the June 2022 issue.

2.0. Preface

In this chapter, I describe research carried out as a collaboration between the groups of Dr. Fredrik Lanner at the Karolinska Institutet and Dr. Gioele La Manno at the EPFL. The findings described here are mostly adapted from the postprint version of a research article entitled “Molecular profiling of stem cell-derived retinal pigment epithelial cell differentiation established for clinical translation.” This work was published in *Stem Cell Reports* in June 2022. I am a co-first author of this publication along with Dr. Sandra Petrus-Reurer (a former member of Dr. Lanner’s group), and we collaborated closely throughout the project. Some additional work presented in Chapter 2.8 describes unpublished results from a follow-up study led by Laura Baqué Vidal (a current member of Dr. Lanner’s group).

This work is the product of combined scientific expertise from two labs: Dr. Lanner’s group has an experimental background in stem cell biology, embryology, and retinal development, whereas our group has a computational background in single-cell sequencing technologies, data analysis, and neurodevelopment. For this project, we also collaborated with Novo Nordisk, which is working to translate the described *in vitro* differentiation protocols into an actionable therapy to treat age-related macular degeneration.

Here, I performed all computational analyses, interpreted the results, designed the figures, and wrote and revised the manuscript. The experimental work was largely performed at the Karolinska Institutet or Novo Nordisk, mostly by Sandra and Laura. However, I helped coordinate experimental work based on results of my analyses, especially during revisions for some histological stainings performed at EPFL. Our collaboration started in the first days of my PhD, in the summer of 2019, and involved hundreds of hours of (remote) coordination among Fredrik, Gioele, Sandra, and myself, to conceptualize the project and transform it into what you read here. All other author contributions are described in Chapter 2.6.2.

2.1. Synopsis

Human embryonic stem cell-derived retinal pigment epithelial cells (hESC-RPE) are a promising cell source to treat age-related macular degeneration (AMD). Despite several ongoing clinical studies, a detailed mapping of transient cellular states during in vitro differentiation has not been performed. Here, we conduct single-cell transcriptomic profiling of an hESC-RPE differentiation protocol that has been developed for clinical use. Differentiation progressed through a culture diversification recapitulating early embryonic development, whereby cells rapidly acquired a rostral embryo patterning signature before converging toward the RPE lineage. At intermediate steps, we identified and examined the potency of an NCAM1+ retinal progenitor population and showed the ability of the protocol to suppress non-

RPE fates. We demonstrated that the method produces a pure RPE pool capable of maturing further after subretinal transplantation in a large-eyed animal model. Our evaluation of hESC-RPE differentiation supports the development of safe and efficient pluripotent stem cell-based therapies for AMD.

2.2. Introduction

The eye, by virtue of its accessibility and relatively isolated anatomical location, has emerged as a promising organ for gene and cell-based therapies to treat neurodegenerative diseases. A pathology that is particularly promising to tackle with these approaches is age-related macular degeneration (AMD), a major cause of severe vision loss affecting more than 180 million people globally (Gehrs et al., 2006). The dry form of the disease, for which no treatment is available, affects 80-90% of advanced patients and is characterized by well-demarcated areas of retinal pigment epithelium (RPE) loss and outer retinal degeneration (Ambati et al., 2003; Sunness, 1999). Human pluripotent stem cell (hPSC) derived RPE cells are thus of high interest for the development of cell replacement treatment options to halt disease progression, as currently being tested in several clinical trials (da Cruz et al., 2018; Kashani et al., 2018; Mandai et al., 2017; Nagiel et al., 2015; Song et al., 2015).

Notable efforts have been made towards developing strategies to ensure high purity RPE products using cell surface markers (Choudhary & Whiting, 2016; A. Plaza Reyes et al., 2020). However, the focus on final product composition has often overshadowed the characterization of intermediate stages appearing before a final steady state is reached. This gap has also been determined by the difficulty of deploying techniques that could systematically distinguish mixed phenotypes and off-target effects from cell heterogeneity and that would allow for a quantitative comparison to physiological references. With this perspective, the availability of single-cell RNA sequencing (scRNA-seq) represents a compelling opportunity.

scRNA-seq can systematically phenotype cell populations produced by differentiation protocols, and its genome-wide readout is crucial to explore the unfolding of *in vitro* differentiation (Kulkarni et al., 2019; Kumar et al., 2017; Lederer & La Manno, 2020). For example, scRNA-seq can determine whether cells follow developmental or non-canonical paths to maturation (Cuomo et al., 2020; McCracken et al., 2020; Veres et al., 2019). In fact, performing an unbiased analysis of cell pools at intermediate stages might expose interesting relations between *in vitro* and *in vivo* processes and help to correctly identify potential risk sources for clinical translation (Begbie, 2013a; Grove & Monuki, 2020; La Manno et al., 2016). Comprehensive single-cell atlases of embryonic and postnatal neurodevelopment are

fundamental to assist in the evaluation of gene expression profiles measured *in vitro* (La Manno et al., 2021; Zeisel et al., 2018). Recent work has sought to decompose cellular heterogeneity of the embryonic and postnatal eye with scRNA-seq, but the similarity between the transient cell states arising in development and hPSC-derived intermediates on the route to RPE lineage has not yet been evaluated (Collin et al., 2019; Cowan et al., 2020; Hu et al., 2019; Lidgerwood et al., 2021; Lo Giudice et al., 2019; Lukowski et al., 2019; Mao et al., 2019; Menon et al., 2019; Shekhar et al., 2016; Sridhar et al., 2020; Voigt et al., 2019). Importantly, both a reference-driven and an unbiased evaluation of the hPSC-RPE cell pool composition at different time points of the protocol in multiple cell lines are critical checkpoints for ensuring a safe and efficient RPE-based replacement therapy.

In this study, we performed scRNA-seq analyses during human embryonic stem cell RPE (hESC-RPE) differentiation using a directed and defined protocol established for clinical translation (A. Plaza Reyes et al., 2020). We demonstrate that the derived cells follow embryonic retinal specification, reaching a mature, pigmented RPE phenotype and even undergoing further maturation towards an adult-like state upon subretinal transplantation into the albino rabbit eye. These findings provide valuable insight into the developmental program of hESC-RPE differentiation and illustrate the required high quality of the derived cells to be used as a future certified clinical product.

2.3. Results

2.3.1 Human embryonic stem cells traverse gene expression space and sequentially mature into retinal pigment epithelium

To examine the differentiation process by which hESC-RPE is generated, we performed time course scRNA-seq (A. Plaza Reyes et al., 2020; Alvaro Plaza Reyes et al., 2016). hESCs were differentiated on human recombinant laminin-521 (hrLN-521) or -511 (iMatrix-511) using NutriStem hPSC XF medium to promote neuroepithelium induction. Activin A was provided on day 6 as a substitute for mesenchymal signaling to induce RPE fate (Cvekl & Wang, 2009; S. Fuhrmann et al., 2000; Fujimura, 2016). Cells were dissociated and replated on day 30 without Activin A, and maturation was completed by day 60 (**Fig. 2.1A**). We profiled differentiation of one research grade cell line (HS980) and two cell lines established for clinical use (KARO1 and E1C3) at six time points during the differentiation protocol (D7, D14, D30, D38, D45, and D60; **Table 2.1**). Morphological evaluation of brightfield images using quantitative cobblestone junction scores confirmed that changes in cell shape and size corresponded with the intended differentiation, as cells progressively assumed a tighter cobblestone monolayer of pigmented cells until D60 (Joshi et al., 2016) (**Fig. 2.1B, S2.1A-B**).

We next performed a global assessment of 26,615 single-cell transcriptomes to verify overall population progression throughout RPE differentiation. Visualizing how cells traversed a reduced gene expression space using the principal components showed that, with time, cells gradually progressed from the pluripotent state towards a mature RPE identity (**Fig. 2.1C, S2.1C-D).** Using 31 marker genes for pluripotent, retinal progenitor, and RPE identities, gene signature scores detected a loss of the pluripotency signature (23.7-fold decrease in signature score from hESC to D30), an increased progenitor status at intermediate days (2.7-fold increase from hESC to D30), and a rise of mature RPE upon protocol conclusion (1.8-fold increase from D38 to D60) (**Fig. 2.1D-E**). Temporal assessment of gene expression confirmed a coherent sequence of expression waves, with pluripotency genes *(POU5F1, LIN28A, SOX2)* leading and being downregulated in favor of progenitor genes (*RAX*, *PAX6*, *VSX2*), eventually trailed by early *(MITF, TYRP1, PMEL, TMEFF2)*, intermediate *(TYR, RLBP1)* and late *(RPE65, BEST1, RGR)* RPE maturation genes (Brandl et al., 2014; Schmitt et al., 2009; Sparrow et al., 2010) (**Fig. 2.1E**). Cells from all three lines were uniformly distributed along the differentiation time course, demonstrating robustness and reproducibility of the protocol (**Fig. 2.1F**). These findings indicated that monolayer differentiation drives the cell pool towards RPE maturation through a path broadly consistent with the developmental process intended to be recapitulated *in vitro*.

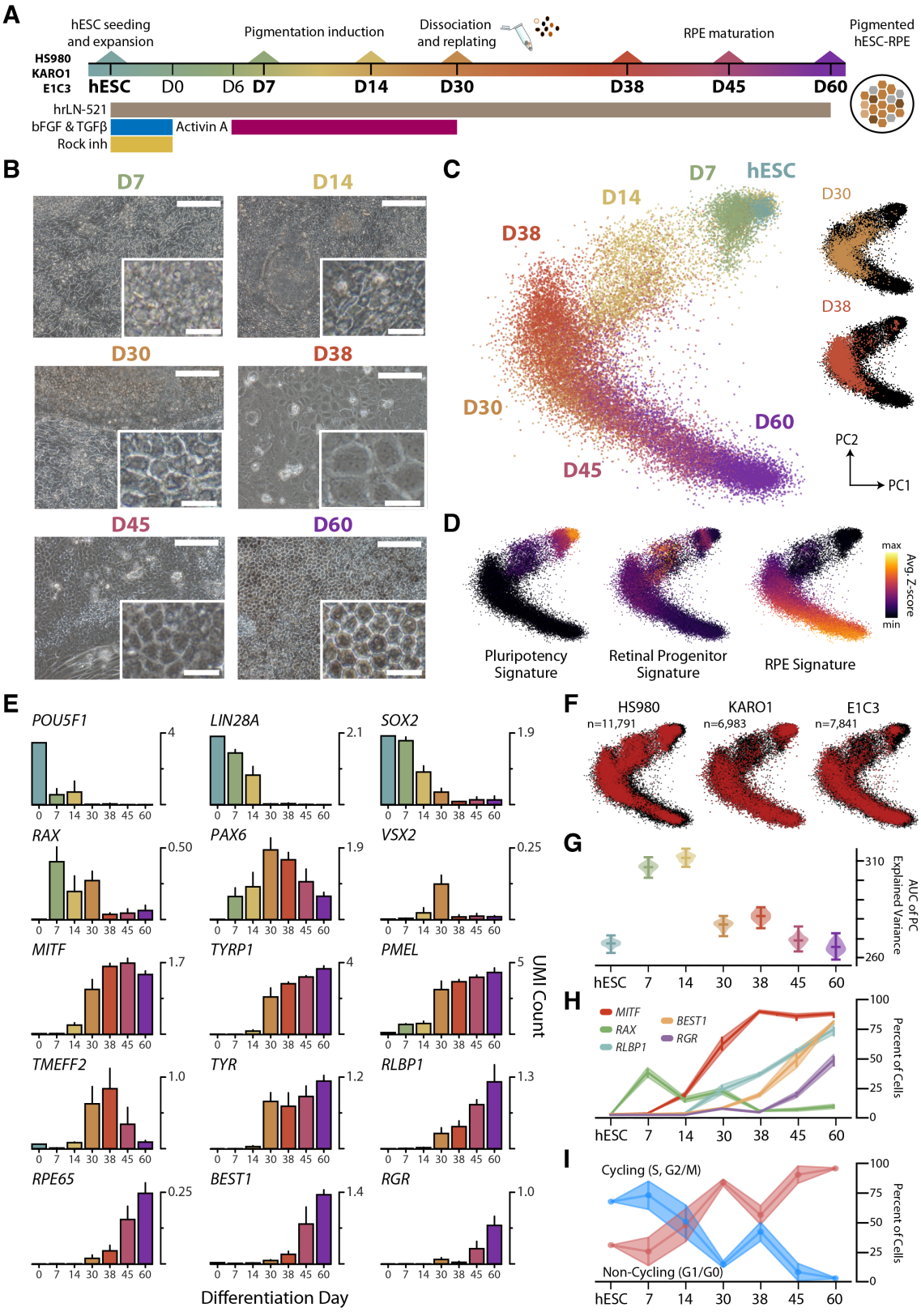


Figure 2.1. Global scRNA-seq characterization of hESC-RPE differentiation trajectory. (**A**) Experimental setup where scRNA-seq was performed at the seven bolded time points in three cell lines: HS980, KARO1, and E1C3. (**B**) Brightfield images during HS980 differentiation. Scale bars: 100μm; inset 20μm (**C**) Principal component (PC) representation of 26,615 single cells across three lines using 2,000 cv-mean enriched genes. (**D**) Signature scores for pluripotency, retinal progenitors and RPE cells.

(**E**) Average normalized gene expression of pluripotent, retinal progenitor, and RPE markers in scRNA-seq data. Error bars represent standard deviation of the mean across three lines, except for the hESC time point. (**F**) PC plot colored by cell line in red. (**G**) Cumulative explained variance curve for each time point and all lines, applied to estimate how much variance accumulates over sets of correlated genes (biological-driven variability), as opposed to uniformly across genes (white noise). (**H**) Percentage of cells positive for retinal marker genes at each time point. (**I**) scRNA-seq based cell cycle phase assignment. Cycling: S and G2/M; Non-Cycling: G1/G0. Intervals in H and I represent the 95% confidence intervals. See also Figure S2.1.

2.3.2. Heterogeneity analysis reveals changes in cell diversity during differentiation

Interestingly, we observed deviations from a uniform progression towards RPE. Cells at D30 appeared more morphologically differentiated towards RPE than those at D38, likely a response to dissociating and replating (**Fig. 2.1C, S2.1A**). A subset of intermediate cells also did not exhibit a strong signature for any of the three identities considered (**Fig. 2.1D**). This suggested a more complex and nonlinear differentiation process than anticipated as well as the presence of additional cell types not captured by our global analyses.

Intrigued, we sought to harness the full phenotyping potential of scRNA-seq and achieve an in-depth description of heterogeneity at all stages. We first calculated how much variance accumulated in correlated gene modules as opposed to uncorrelated genes and interpreted this quantity as a measure of biological heterogeneity (**Methods 2.5; Fig. S2.1F**). Analysis of these values computed on all studied cell lines revealed that hESCs and D60 cells harbored a lower heterogeneity compared to intermediate time points, particularly D7 and D14 (**Fig. 2.1G**). While a large decrease in heterogeneity was detected from D14 to D30, suggesting an initial convergence towards RPE fate, we observed a slight increase from D30 to D38, hinting at an effect of cell dissociation, replating, or Activin A removal on cell composition. Similarly, a decrease in cobblestone junction scores was observed between D30 and D38, in the HS980 and KARO1 but not in the E1C3 lines (**Fig. S2.1A**). Furthermore, initial (hESCs) and final (D60) samples had mutually exclusive and uniform expression of pluripotency and RPE genes (**Fig. 2.1H, S2.1G-I**). This was consistent with proliferation trends: a decreased fraction in cycling cells from hESC to D30, followed by an increase from D30 to D38 and finally a reduction from D38 to D60 (**Fig. 2.1I-J**).

2.3.3. Early differentiation recapitulates cellular diversity of the rostral neural tube and optic vesicle

To determine the identity of cell populations during initial RPE induction (D7, D14, and D30), we obtained enriched genes by population and cross-referenced the literature to

annotate each cluster with a primary (group) and secondary (cluster) categories (**Methods 2.5**). This process revealed a mixture of intermediate cell states resembling those described in the context of rostral neural tube patterning and eye development (Bosze et al., 2020; Sarkar et al., 2020).

To obtain a unified clustering scheme across cell line replicates without neglecting the possibility of line-specific cell types, we first annotated each dataset for HS980, KARO1, and E1C3 samples individually. Then, we performed canonical correlation analysis (CCA) on samples corresponding to the same differentiation day and constructed maps of enriched genes for shared cell types (**Methods 2.5**) (Butler et al., 2018). At D7 and D14, we identified seven groups based on their expression profiles: Pluripotent-like (Pluri), Endodermal-like (Endoderm), Lateral Neural Fold-like (LatNeEp), Pre-Placodal Epithelium-like (PrePlac), Cranial Neural Crest-like (CrNeCr), Mesenchymal (MesCh), and Retinal Progenitor (RetProg) (**Fig. 2.2A**). Several marker genes described in the literature as corresponding to such cell types, including *RAX, DLX5, FOXE3, FOXC1, HAND1, NANOG,* and *SOX17*, supported the annotated cluster identified (**Fig. 2.2B, S2.2A; Table 2.2**) (Bosze et al., 2020; Firulli et al., 2014; Kwon et al., 2010; McLarren et al., 2003; Pan & Thomson, 2007; Qu et al., 2008; Seo et al., 2017). This analysis highlighted distinct differences between the three lines, such as that KARO1 retained a pool of pluripotent-like cells in initial time points, HS980 generated more pre-placodal-like cells, and E1C3 initiated an endodermal-like population while also establishing the largest percentage of retinal progenitors (**Fig. 2.2C-D**).

The surprising emergence of different cell types of the anterior ectoderm highlights the interrelatedness of gene expression programs for the eye field, neural crest, and other sensory tissues during early embryonic development. Some secondary clusters, particularly in the HS980 line, matched remarkably well with specific neural tube regions, expressing a combination of enriched markers for eye field *(RAX, SIX6, LHX2)*, telencephalic neural fold *(DLX5*, *DLX6)*, lens placodes *(FOXE3*, *PAX6*, *ALDH1A1)*, cranial neural crest *(FOXC2, VGLL2, PITX1)*, inner ear placodes *(OTOGL, VGLL2, CYP26C1)*, the anterior neural ridge organizer *(FGF8, SP8, and FOXG1)*, and mesenchyme *(GABRP, HAND1, COL1A1)* (Cajal et al., 2012; J. Chen et al., 2017; Cohen-Salmon et al., 1997; Crespo-Enriquez et al., 2012; Gitton et al., 2011; Kasberg et al., 2013; Kumamoto & Hanashima, 2017; Seo et al., 2017; Soldatov et al., 2019; Tahayato et al., 2003) (**Fig. 2.2E; Table 2.2**).

The observation of mesenchyme in all three cell lines is interesting because periocular mesenchyme expresses inductive signals *in vivo* that promote RPE fate, which has been carried out by Activin A in the present protocol (Bosze et al., 2020). Eye field-like cells (RetProg) detected across the differentiation possessed distinct gene expression programs,

suggestive of varying degrees of progression towards RPE. RetProg clusters expressed a repertoire of known markers, including *OTX2* and *LHX2*, which are jointly necessary for activation of the transcription factor *MITF*. At D14, these two genes were co-expressed in RetProg clusters alongside MITF-activated genes *PMEL*, *SERPINF1*, *TYRP1* and *DCT* (**Fig. S2.2B-D**). Consistent with their classification as progenitors, cells displayed a stark cell proliferation signature (S and G2/M phases) (**Fig. S2.2E**). CCA of the D7 and D14 populations further captured a "pseudospatial" axis of variation, with cells transitioning along a mediolateral molecular profile (**Fig. 2.2F-G**).

We next integrated D7 and D14 cells from all three lines onto a shared feature space to factor out time-dependent differences (**Fig. S2.2F**). Over time, we observed an increase in cells assigned as pre-placodal (PrePlac; 5.24% to 34.59% of cells) and a decrease in both inner ear-like cranial neural crest (CrNeCr; 30.36% to 9.30%) and lateral neuroepithelial (LatNeEp; 16.84% to 5.01%) cells. The fraction of retinal progenitors remained more stable (RetProg; 33.48% to 24.65%) (**Fig. S2.2F-G**). These analyses revealed that this heterogeneity at both stages recapitulates the molecular profile of rostral embryonic territories patterned to specify sensory organs, such as lens, olfactory, and otic placodes (Begbie, 2013b) (**Fig. 2.2G**).

Conversely, between 79% and 96% of cells at D30, depending on the cell line, were categorized into retinal progenitor or RPE stages (RetProg, EMT-RPE, EarlyRPE, MidRPE, LateRPE). Other observed cell types included Lateral Neural Fold-like (LatNeEp; 1.1-4.0%), Neuronal (Neuronal; 0.2-4.2%), Mesenchyme-like cluster 1 (MesCh1; 0.4-5.6%), Mesenchyme-like cluster 2 (MesCh2; 0.1-3.7%), Neural Retina (NrlRet; 0.3-2.2%), and Floor Plate (FloorPlate; 0.5-2.4%) (**Fig. 2.2H, S2.2H**). A pseudotemporal trajectory of retinal maturation largely characterized these cells (**Methods 2.5; Fig. 2.2I**). Expression along this pseudotime confirmed a loss of progenitor status *(SOX2, RAX, VSX2, LHX2)*, followed by an increase in RPE differentiation *(MITF, TYRP1, PMEL)* and, later, of advanced RPE maturation markers *(TYR, RLBP1, RPE65, BEST1, RGR)* (**Fig. 2.2J**). Transcription factor network analysis of D30 HS980 cells with SCENIC further confirmed the activity of regulons involving factors *SOX2*, *RAX*, *VSX2*, *OTX2*, and *MITF* with anticipated gene targets (Aibar et al., 2017) (**Fig. S2.2I**). In fact, MITF cooperates with OTX2 to transactivate RPE pigmentation genes and downregulate progenitor genes (Martínez-Morales et al., 2003; Yun et al., 2009).

Our observations indicate that a sequential stepwise differentiation model is inadequate to explain the observed cell population dynamics; instead, the data suggests a "divergence-convergence" model, with an initial expansion of cellular diversity, later dampened to favor the promotion of the RPE differentiation program (**Fig. S2.2J; Discussion 2.4**).

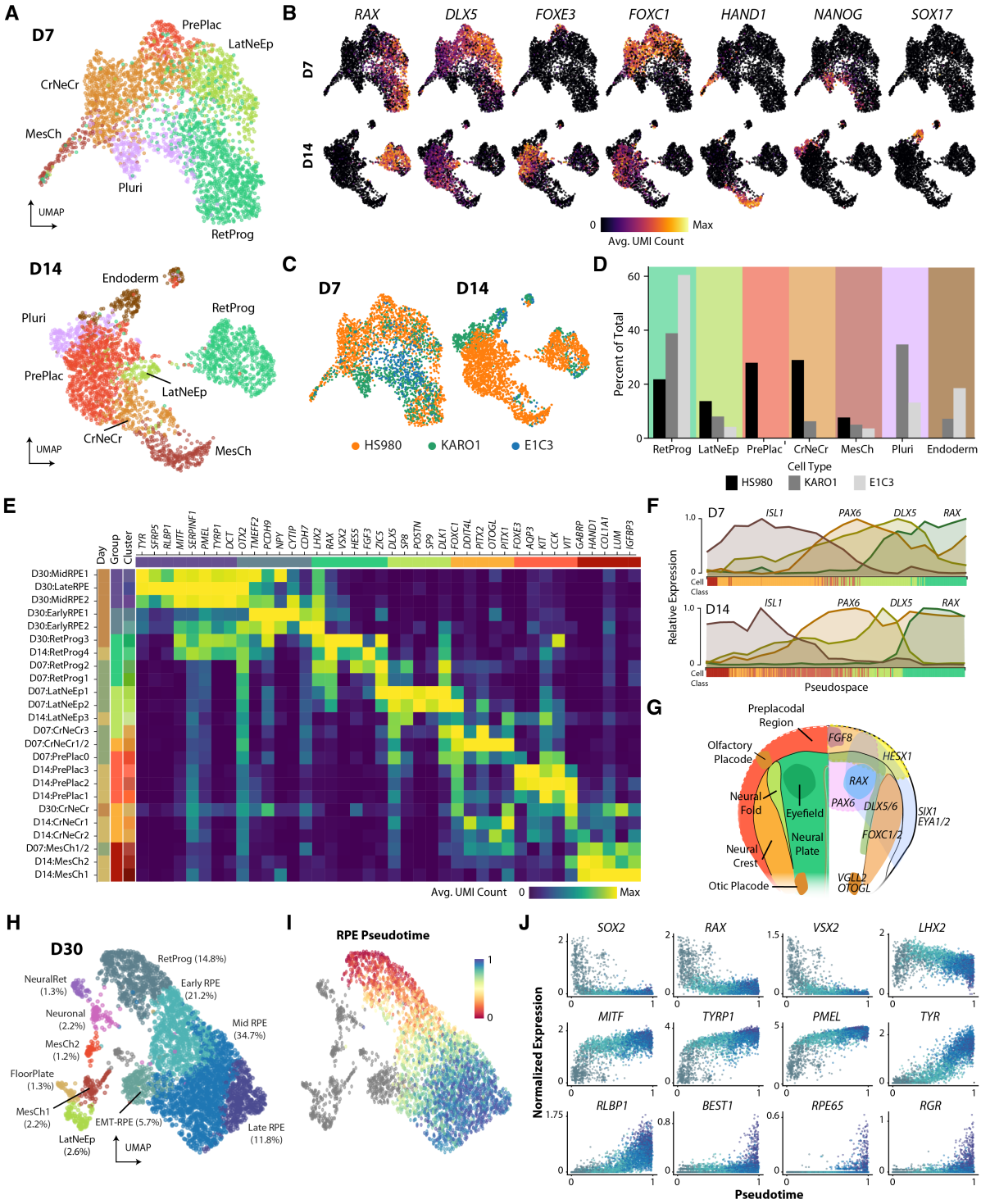


Figure 2.2. Evaluation of diverse neuroepithelial cell type derivatives in early differentiation. (**A**) UMAP at differentiation D7 and D14 in three lines. Cells were grouped into Retinal Progenitor (RetProg), Lateral Neural Fold-like (LatNeEp), Pre-Placodal-like (PrePlac), Cranial Neural Crest-like (CrNeCr), Mesenchyme (MesCh), Pluripotent (Pluri), and Endoderm-like (Endo) clusters. (**B**) Normalized gene expression of marker genes *RAX* (RetProg), *DLX5* (LatNeEp), *FOXE3* (PrePlac), *FOXC1* (CrNeCr), *HAND1* (MesCh), *NANOG* (Pluri), and *SOX17* (Endo). (**C**) UMAPs in (A) colored by cell line. (**D**) Cell type composition in each line at D7 and D14. (**E**) Enriched gene expression heatmap for HS980 cell types. (**F**) Plots showing relative expression of neural tube patterning markers in D7 (top) and D14

(bottom) cells across pseudospace in HS980. (**G**) Schematic of the patterned anterior neural plate at the neurulation stage. Left: putative location of the cell types corresponding to identified clusters. Right: schematic of genes patterning the rostral embryo. (**H**) UMAP at differentiation D30 three lines. (**I**) Pseudotime trajectory of D30 RPE and RetProg cells (82.5% of total at D30). (**J**) Progenitor (*SOX2, RAX, VSX2, LHX2*), early (*MITF, TYRP1, PMEL*), mid (*TYR, RLBP1*) and late (*RPE65, BEST1, and TTR*) gene expression along pseudotime. See also Figure S2.2.

2.3.4. 2D monolayer differentiation is faster and more directed than 3D embryoid body differentiation

The molecular patterning of 2D monolayer cultures during early RPE differentiation hints at an intriguing self-organization process taking place despite the lack of spatially directed cues or a 3D structure. The spontaneous generation of alternative cell fates has been described in studies investigating intermediate stages of other differentiation protocols (Cuomo et al., 2020; La Manno et al., 2016; Lin et al., 2021). Furthermore, previous reports have suggested that a highly spatially organized system is not necessarily the most molecularly patterned (Quadrato et al., 2017; Velasco et al., 2019). To clarify how the initial heterogeneity detected in our monolayer differentiation relates to a 3 dimensional (3D) protocol that allow cells to organize spatially, we compared to an Embryoid Body (EB) differentiation protocol (Alvaro Plaza Reyes et al., 2016).

We performed hESC-RPE differentiation in the EB setting for 30 days, followed by scRNA-seq (2,851 cells) at D7, D14, and D28. At D7, EBs displayed a mostly uniform patterning with early progenitor and pluripotency markers (*SOX2*, *PAX6, SALL4, LIN28A;* **Fig. S2.3A**). Conversely, D14 EBs showed the emergence of cells corresponding to the three primary brain vesicles: we could distinguish prosencephalon-, mesencephalon-, and rhombencephalon-like (Fore-, Mid-, Hindbrain) clusters (**Fig. 2.3A, S2.3B**). The presence of the observed patterning-related heterogeneity was confirmed by a signature score analysis based on markers of different brain and rostro-caudal neural tube regions (**Fig. 2.3B**). Interestingly, we detected no traces of the more specific mediolateral-patterning signatures observed in early stages of the 2D protocol (**Fig. 2.3C**). Consistently, midbrain and hindbrain gene expression signatures were not detected in the 2D cultures, and a greater fraction of retinal progenitors were observed in the 2D context (24.7% 2D cells compared to 1.6% 3D cells at D14) (**Fig. 2.3D, S2.3C**).

At D28, the 3D EB culture continued to harbor more diversity than the 2D monolayer (**Fig. 2.3D**). While 2D cultures at D30 were largely defined by various stages of RPE maturation, in the 3D EB culture other retinal and brain-related cell types were present, including populations of the neural retina (Progenitor-like, Horizontal cell-like, and Amacrine-like) as well as caudal neuroblasts and glial populations (Roof Plate, Neuroblast 1 and 2,

Posterior Neuronal, Choroid Plexus, Schwann Precursor Cells) (**Fig. 2.3E, S2.3D**) (Brodie-Kommit et al., 2021; La Manno et al., 2021; Lu et al., 2020). This evidence, in conjunction with the presence of distinct *TBX2*+ dorsal optic cup-like and *VAX2*+ ventral optic cup-/stalk-like progenitor populations, strongly suggests that a wider set of morphogenetic events, not all related to RPE differentiation, are recapitulated during the EB protocol (Behesti et al., 2006; Bosze et al., 2020).

Finally, the transcriptomic correspondence between cells in the 2D and 3D protocols was evaluated more directly by projecting 2D-cultured cells onto the EB D28 embedding. The large majority of 2D cells mapped to RetProg and RPE clusters rather than non-retinal cell types, illustrating that RPE cells comprise a much larger fraction of the monolayer cultures (73.4% cells) than of the EB cultures (10.2% cells) (**Fig. 2.3E-F**).

In summary, the 3D EB hESC-RPE differentiation protocol produces cell identities corresponding to a broader neural origin and drives differentiation of multiple retinal lineages in parallel with RPEs. Conversely, the 2D protocol tends to specify more quickly and narrowly to rostral embryonic cell identities, further funneled to a RPE fate around D30, thus supporting a divergence-convergence model.

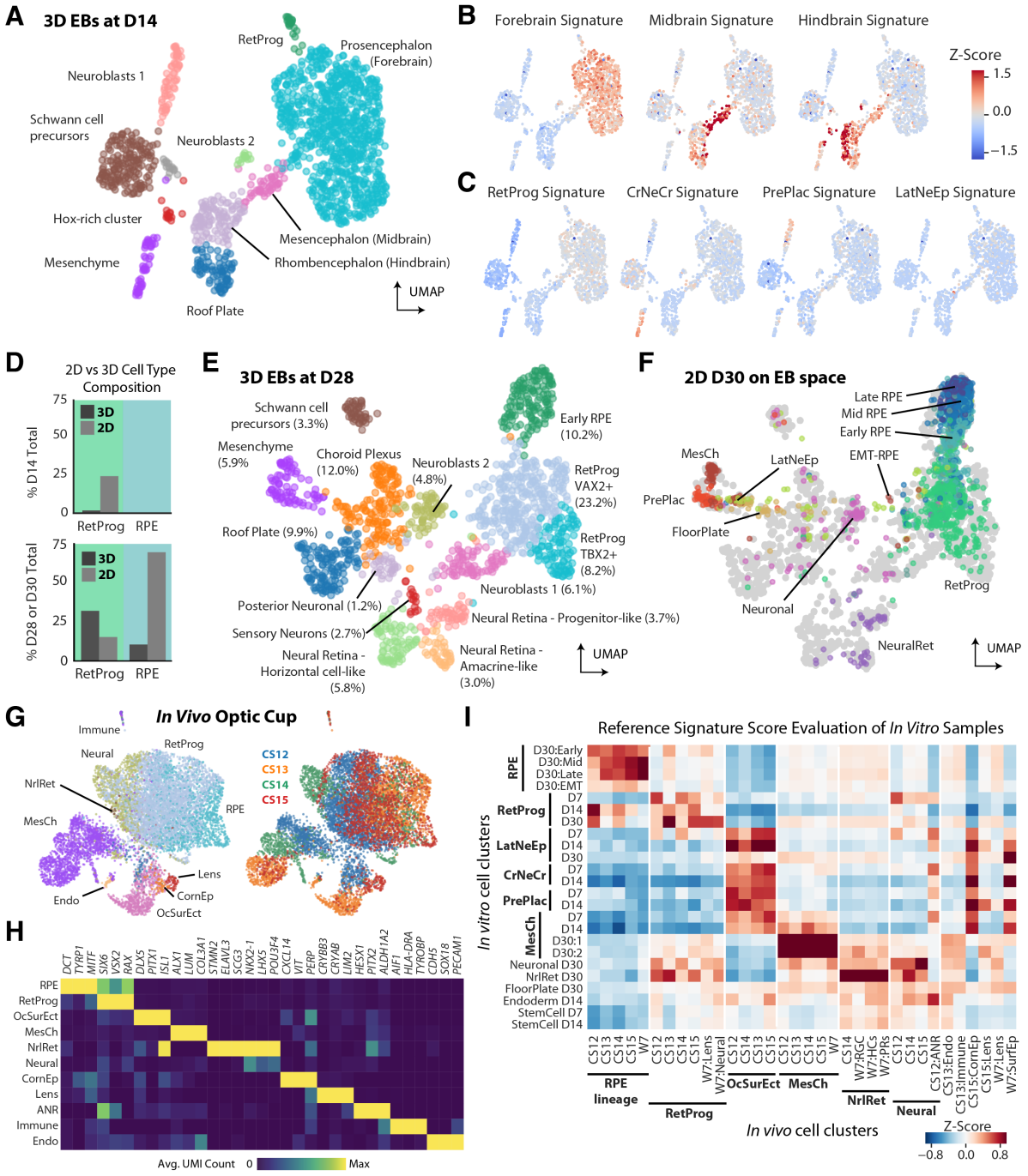


Figure 2.3. Comparative analysis of RPE induction between hESC-RPE, 3D EB differentiation, and embryonic eye. (A) UMAP of 3D EB cultures at D14 (HS980 line). **(B-C)** Signature scores for brain regions (B) and neural tube cell types (C) visualized on the EB D14 UMAP. **(D)** Bar plots comparing cell type compositions in 2D and 3D cultures at D14 (top) and D28/30 (bottom). **(E)** UMAP of 3D EB cultures at D28 colored by cell type. **(F)** Projection of 2D D30 cells from all three cell lines onto the UMAP from (E) using pairwise correlation distances, colored by annotated cell type (see Supplemental Experimental Procedures, cf. Figure 2.2H). Cells in gray are those from (E). **(G)** UMAP of human embryonic optic cup cells at Carnegie Stages 12, 13, 14, and 15 (week 5), colored by cell type or stage. **(H)** Heatmap of enriched gene expression by cell type across all samples in (G). (I) Signature scores of *in vitro* cell clusters at D7, D14, and D30 illustrating the correspondence to *in vivo* clusters from (G). Signature scores were obtained using the top 30 genes of the respective *in vivo* reference population. See also Figure S2.3.

2.3.5. *In vitro* differentiation and eye development exhibit similarities in cellular composition

We reasoned next that embryonic references could validate our model and evaluate how faithfully *in vitro* phenotypes match their *in vivo* counterparts. Thus, we performed scRNA-seq on four human embryonic optic vesicles from Carnegie Stages (CS) 12, 13, 14, and 15 (approximately 30-, 32-, 33-, and 36-days post-conception; 9,409 cells) as well as two eyes at Carnegie Stage 20 (7.5 weeks post-conception; 2,742 cells). Early stages contained patterned cell types corresponding to optic vesicle and surrounding tissues, including retinal progenitors (RetProg), retinal pigment epithelium (RPE), neural retina (NrlRet), ocular surface ectoderm (OcSurEct), lens (Lens), mesenchyme (MesCh), other neural (Neural), immune (Immune), the anterior neural ridge (ANR), and endothelial cells (Endo); these populations were mostly well distributed across developmental stages (**Fig. 2.3G-H**). Retinal tissues were already more clearly differentiated into embryonic RPE (eRPE), neural retinal, and optic stalk sub-populations by CS13 (W5) than in the RPE-focused progenitors detected *in vitro* (**Fig. S2.3E, cf. Fig. 2.2E, S2.2A; Table 2.3**). The CS20 (W7.5) samples captured a more diverse representation of cell types surrounding the eye, including proliferating progenitors, RPE, lens, intermediate retinal ganglion cells, and neural crest-derived mesenchyme (**Fig. S2.3F-G**). Progenitor markers highly expressed in earlier stages were more exclusive to neural (NePr) and lens (LensPr) progenitors by W7.5. Early differentiation genes were detected in embryonic RPE, and a transition towards mature RPE was apparent in the embedding despite an absence of mature RPE markers *RPE65, BEST1*, and *TTR* (**Fig. S2.3H-I**).

To evaluate the resemblance of hESC-RPE clusters to embryonic references, enriched genes were extracted from *in vivo* cell types at all five stages and used to compute signature scores for *in vitro* cell types at D7, D14, and D30 (**Fig. 2.3I**). Scores for the RPE clusters *in vitro* were highest using enriched genes from the *in vivo* RPEs, with signatures of later RPE populations *in vitro* scoring higher against later-stage embryonic RPEs. An inverse correspondence was also observed between *in vivo* RPE signatures and scores for *in vitro* RetProg populations, relating to the gradual maturity of such populations. Furthermore, signatures comprised of enriched genes from the *in vivo* ocular surface ectoderm scored highest for the *in vitro* LatNeEp, CrNeCr, and PrePlac clusters, suggesting that these patterned populations are similar to the ectodermal tissue surrounding the optic cup *in vivo*. Indeed, overlapping gene expression patterns were observed, with OcSurEct expressing markers for *in vitro* tissues such as *DLX5, PITX1,* and *ISL1* (**Fig. 2.3H**). Similar signature score analysis using clusters from a recent atlas of more than 100,000 single cells from *in vivo*

retinas and organoid cultures showed similar patterning, although there were fewer corresponding cell types overall due to the large number of neural retina clusters in the atlas (**Fig. S2.3J**) (Lu et al., 2020). Overall, the molecular profiles of early hESC-RPE populations mirrored those present during development *in vivo*, but with a strong bias towards RPE fate over other retinal cell types.

2.3.6. Cell surface marker NCAM1 defines retinal progenitor cells at D30 of hESC-RPE differentiation

Selective removal of progenitor populations at intermediate stages could be both a route to faster RPE differentiation protocols and a strategy to obtain a cellular source to derive other retinal lineages from. Following the observation of a retinal progenitor cell population in D30 hESC-RPE cultures, we reasoned that reliable retinal progenitor markers would be inversely correlated to genes characterizing more mature cells, such as RPE, as well as to progenitors for other neuroepithelium tissues. We therefore computed a Pearson's correlation coefficient between highly expressed genes at D30 and RPE or neural tube markers (**Fig. S2.4A-B**). Genes with strong anticorrelation to both signatures encoded functionally diverse gene products, including cytoplasmic proteins, transcription factors, secreted molecules, and membrane proteins (**Fig. 2.4A; Table 2.4**). Transcription factors involved in early retina development (*SFRP2, CRABP1*, *RAX*, *SIX6)* ranked among the top genes, along with genes implicated in neural tube (*CPAMD8*, *PKDCC*, *NR2F1*) and lens development (*MARCKS*, *DACH1*, *MAB21L1*) (Imuta et al., 2009; Yamada et al., 2003; Zhou et al., 2010).

Interestingly, cell surface markers *CDH2*, *CPAMD8*, and *NCAM1* were among the most prominent progenitor markers at D30. We were intrigued particularly by *NCAM1* (surface antigen CD56), which was identified previously in a screen for early RPE markers and whose expression was shown to be anticorrelated with pigmentation (A. Plaza Reyes et al., 2020). We found that NCAM1 staining areas coincided with rosette structures lacking pigmentation, and both scRNA-seq and protein staining revealed that NCAM1 was co-expressed with progenitor (VSX2 and RAX) and proliferative (Ki67) genes at D30 (**Fig. 2.4B, S2.4C-E**).

To functionally examine the nature of NCAM1-positive progenitors and whether they hold potential to generate RPE cells, we devised a sorting strategy to isolate this population using NCAM1 and CD140b (PDGFRB), an RPE cell marker (A. Plaza Reyes et al., 2020). By combining the two markers, we separated cells into either a putative retinal progenitor stage (24% cells, CD140b^{low}NCAM1^{high}, henceforth NCAM1-High) or a more mature RPE stage (56% cells, CD140b^{high}NCAM1^{low}, henceforth CD140b-High) (**Fig. 2.4C**). Consistently, pigmentation was evident in the CD140b-High population cell pellets, whereas the NCAM1-

High population lacked pigmentation, suggesting different intrinsic potentials and maturation statuses (**Fig. 2.4D**).

To first characterize the sorted NCAM1-High and CD140b-High populations, we performed scRNA-seq analysis directly after sorting (NCAM1-High: 734 cells; CD140b-High: 1,486 cells). Following integration with unsorted D30 cells (1,852 HS980 cells; see **Fig. 2.2H**), NCAM1-High cells predominantly corresponded to RetProg (64.7%) and EarlyRPE (20.6%), whereas CD140b-High cells were mostly of EarlyRPE (28.0%), MidRPE (42.9%) and LateRPE (17.6%) profiles; unsorted D30 cells were more evenly distributed among cell type states (**Fig. 2.4E-F, S2.4F-G**). Consequently, NCAM1-High cells were enriched for the expression of progenitor markers *(FEZF2, CRB1, SOX2, FGF9, VSX2)* whereas CD140b-High cells were enriched for mature RPE markers *(SFRP5, TTR, SLC35D3, TYR, RLBP1)* (**Fig. 2.4G**).

We then continued RPE differentiation with sorted D30 NCAM1-High and CD140b-High cells for 30 days. Morphological evaluation showed that CD140b-High cells generated a homogeneous hESC-RPE monolayer already at D45, while NCAM1-High cells only yielded a defined RPE morphology at D60 (with cobblestone scores of 6.77E-3 per μm^2 for the CD140b-High population and 5.29E-3 per μm^2 for the NCAM1-High population at D60, correlating with unsorted D30 and D60 hESC-RPE) (**Fig. S2.4H, S2.1A**). The percentage of cells positive for the cell cycle marker Ki67 was higher in the NCAM1-High population, indicative of a proliferative state, and eventually declined upon reaching a RPE phenotype (**Fig. S2.4I**). In fact, VSX2 protein was more abundant in NCAM1-High cells that also co-expressed Ki67 for a longer time under RPE culturing conditions than in CD140b-High cells (**Fig. S2.4J**).

We assessed retinal progenitor *(SIX6, VSX2, RAX, PAX6) and RPE (MITF, BEST1, RPE65, TYR)* markers by RT-qPCR to understand the expression dynamics, confirming that NCAM1-High cells expressed higher levels of progenitor genes than CD140b-High cells at the time of sorting. The progenitor genes continued to be expressed after sorting in NCAM-High cultures (declining over time under RPE differentiation conditions), whereas in CD140b-High cultures they were close to absent throughout the protocol. Conversely, CD140b-High cells upregulated mature RPE markers earlier and more rapidly than NCAM1-High cells (**Fig. 2.4H, S2.4K**).

Establishment of an RPE phenotype at D60 by both NCAM1-High and CD140b-High populations was confirmed by scRNA-seq of sorted and unsorted populations (NCAM1-High: 1,106 cells; CD140b-High: 987 cells; unsorted: 975 cells from an additional replicate). Both contained large proportions of MidRPE and LateRPE (65.0% in NCAM1-High, 97.1% in CD140b-High, and 88.4% in unsorted D60) and similar expression levels of early and mature RPE markers (**Fig. 2.4I-K, S2.4L-M**). The appearance of a cobblestone morphology and

pigmented cultures in addition to co-expression of CD140b and BEST1 proteins was also confirmed in both sorted populations (**Fig. 2.4L**).

To evaluate the functional relevance of those changes and compare the degree of differentiation between the two sorted populations, we assessed pigment epithelium-derived factor (PEDF) secretion and transepithelial resistance (TEER) upon protocol completion (D60), finding that CD140b-High-derived cells secreted significantly higher apical levels of PEDF than the unsorted and NCAM1-High-derived cells, both at similar standard levels (1000-2000 ng/mL) (**Fig. 2.4M**). TEER levels displayed by CD140b-High-derived cells were higher compared to unsorted and NCAM1-High populations, whose levels were comparable to D60 hESC-RPE cells (400-800 $\Omega*cm^2$) (A. Plaza Reyes et al., 2020; Alvaro Plaza Reyes et al., 2016) (**Fig. 2.4N**). These results show that CD140b selects for more mature pigmented RPE cells at D30, whereas NCAM1 denotes an immature progenitor population with potential to mature into functional RPE cells.

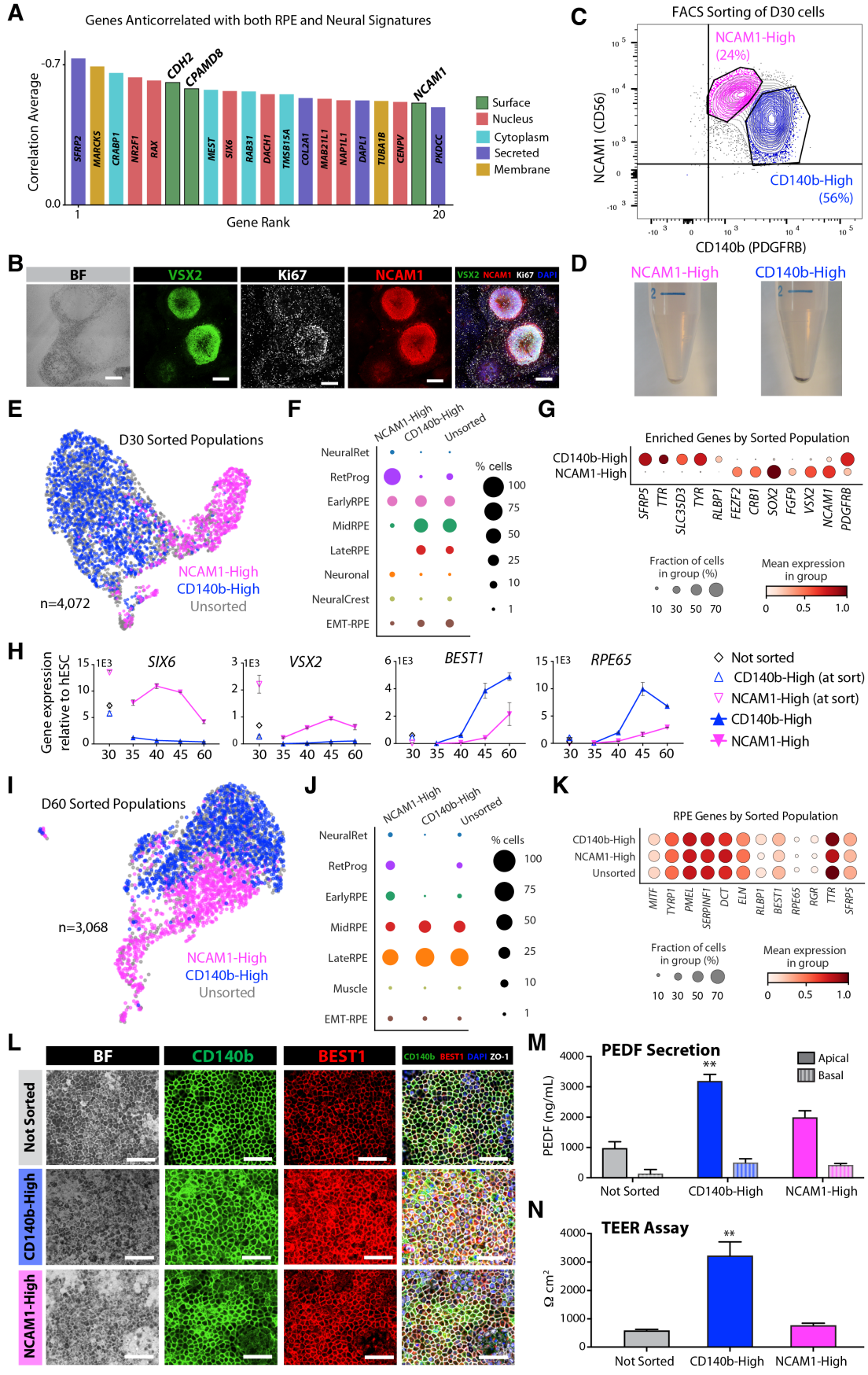


Figure 2.4. Characterization of the NCAM1-High-sorted D30 population. (**A**) Bar graph of top genes from anticorrelation analysis at HS980 D30. Genes with a mean normalized expression <0.5 were

excluded. (**B**) Brightfield and immunofluorescence stainings of D30 cells showing co-expression of VSX2, NCAM1, and Ki67 markers. Scale bars: 200μm. (**C**) Representative FACS plot of NCAM1-CD140b sorting to distinguish distinct populations at D30. (**D**) Post-sort pellets of CD140b-High and NCAM1-High cells. (**E**) UMAP of NCAM1-High (pink), CD140b-High (blue), and unsorted (gray) D30 cells after CCA integration. (**F**) Dot plot illustrating the proportion of cells corresponding to each identified cell type in scRNA-seq samples from (E). (**G**) Dot plot of selected progenitor (*FEZF2*, *CRB1*, *SOX2*, *FGF9*, *VSX2*) and RPE (*SFRP5*, *TTR*, *SLC35D3*, *TYR*, *RLBP1*) genes enriched in the sorted samples. (**H**) RT-qPCR of retinal progenitor (*SIX6*, *VSX2*) and RPE (*BEST1*, *RPE65*) marker genes in populations from (**E**) at the moment of sort and at post-sort D30, D35, D40, D45, and D60. (**I**) UMAP of NCAM1-High (pink), CD140b-High (blue), and unsorted (gray) D60 cells. (**J**) Dot plot illustrating the proportion of cells corresponding to each identified cell type in scRNA-seq samples from (I). (**K**) Dot plots of early (*MITF, TYRP1, PMEL, SERPINF1, DCT, ELN*) and late (*RLBP1*, *BEST1*, *RPE65*, *RGR*, *TTR*, *SFRP5*) RPE genes in the LateRPE cell clusters from each sorted sample. (**L**) Brightfield and immunofluorescence stainings of unsorted, CD140b-High and NCAM1-High populations 30 days after sorting (D60) showing co-expression of CD140b, BEST1 and ZO-1 markers. Scale bars: 100μm. (**M-N**) PEDF secretion (M) and TEER measurements (N) of the unsorted, CD140b-High and NCAM1-High populations at D60. (H, M-N). Bars represent mean +/-SEM from three independent experiments. See also Figure S2.3.

2.3.7. NCAM1-High cells can differentiate into alternative retinal cell types

To evaluate the full differentiation potential of NCAM1-High progenitors, the population was sorted at D30 and plated in neuroretinal progenitor-promoting conditions for 40 additional days (Shao et al., 2017) (**Fig. 2.5A**). Interestingly, NCAM1-High cells gave rise to a heterogeneous culture, with a significant portion of cells displaying a distinct non-RPE cell body morphology; this was in contrast to CD140b-High cells, which showed the typical RPE cobblestone profile under the same conditions (**Fig. 2.5B**). Gene enrichment analysis of 980 single cells yielded a variety of molecularly-distinct populations, of which only 12% were RPE, suggesting NCAM1-High cells at D30 represent an uncommitted progenitor with potential beyond RPE whereas CD140b-High captures lineage-committed RPE cells (**Fig. 2.5C**).

We then performed CCA integration with the CS20 (W7.5) embryonic eyes to systematically compare NCAM1-High-derived cells to a developmental reference. The shared low dimensional space emphasized similarities between corresponding clusters, including RPE, progenitor, mesenchymal, lens, surface epithelial and neuronal populations (**Fig. 2.5D-F**). Non-RPE retinal cell types detected included a small lens population co-expressing *LIM2*, *CRYAB*, *PITX3*, and *PROX1*. Genes exclusive to W7.5 embryonic lens *(FOXE3* and *SOX1)* are specific to promoting early lens development, suggesting that NCAM1-High-derived lens cells are in a more mature state (Blixt et al., 2000; Nishiguchi et al., 1998) (**Fig. 2.5G**). A shared epithelial population was also observed, co-expressing markers characteristic of surface epithelium and keratinocytes, which can be found in the cornea (**Fig. 2.5H**).

Moreover, there was an overlap between W7.5 retinal ganglion neurons and the NCAM1-High-derived neurons (**Fig. 2.5D-E, cf. Fig. S2.3**). To compare gene expression dynamics of these cells, RNA velocity was computed on each neuronal population using *velocyto*, revealing progression towards a more mature state (La Manno et al., 2018) (**Methods 2.5; Fig. 2.5I**). Pseudotemporal gene expression confirmed a common profile of expression waves, with gradual downregulation of proliferation markers *(TOP2A, MKI67)* followed by upregulation of a neuronal differentiation program *(TAGLN3, STMN2, TUBB2A, DCX, NRXN1)* (**Fig. 2.5J**). However, markers of retinal ganglion development, such as transcription factor *ATOH7* and its downstream targets *POU4F2* and *ISL1*, were only expressed in the embryonic cells (P. Gao et al., 2014) (**Fig. 2.5K**). Other neuronal markers *(EOMES, NEUROD2, NEUROD6, SLA)* were unique to NCAM1-High-derived neurons, implying that NCAM1-High-derived cells are another type of telencephalic neuron (**Fig. 2.5L**). NCAM1-High cells are thus either a mixed pool of retinal and neuroepithelial progenitors capable of forming both cell types and other related retinal lineages, or cells with the capacity to establish all these lineages.

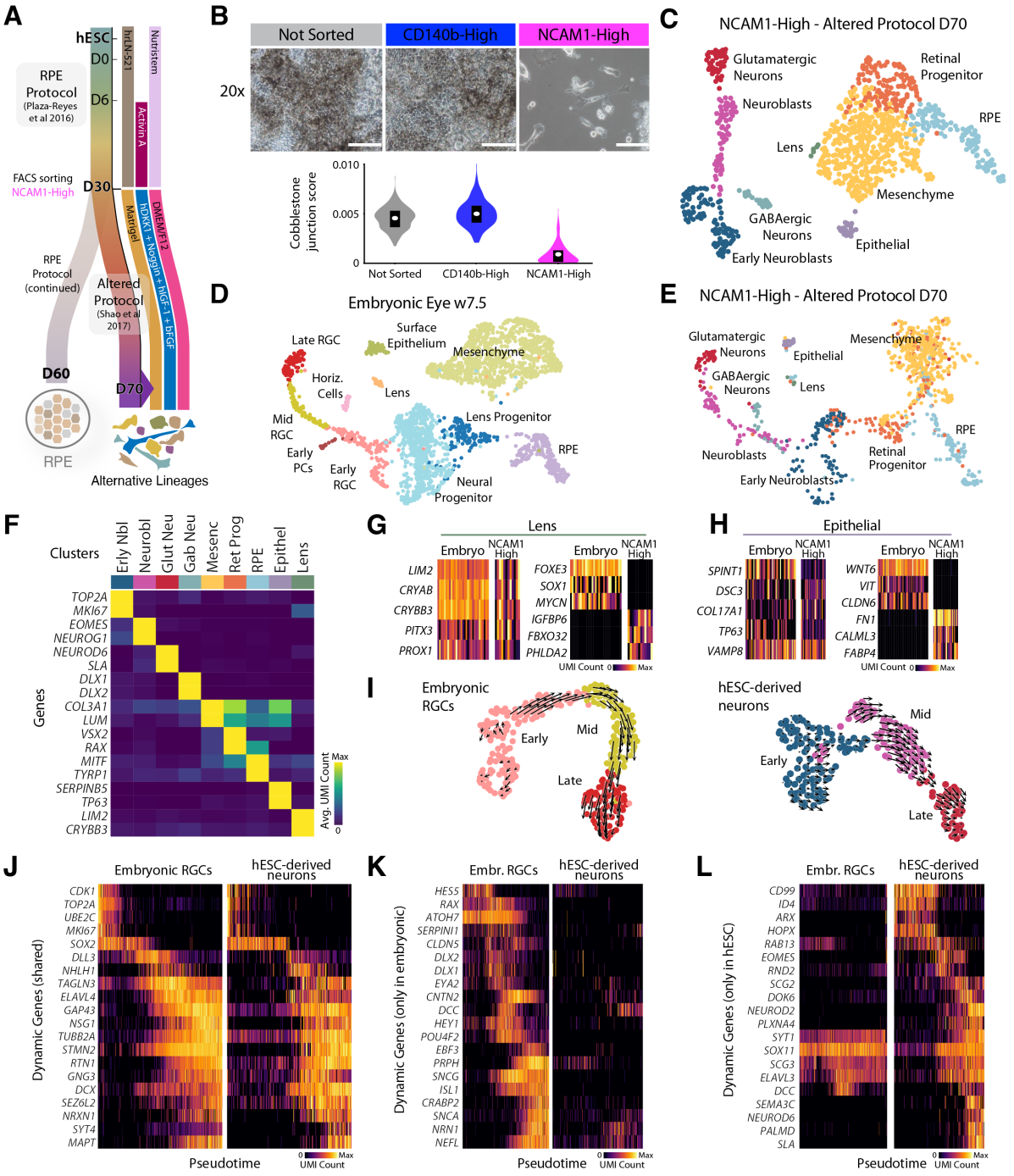


Figure 2.5. Neuroretinal progenitor differentiation of NCAM1-High-sorted cells. (A) Schematic of the neuroretinal progenitor (altered) differentiation protocol. D30 NCAM1-High-sorted cells were sorted and replated on matrigel containing DMEM/F12, hDKK1, Noggin, hIGF-1, and bFGF until scRNA-seq at D70. (**B**) Brightfield images and cobblestone junction scores of sorted and unsorted populations at D70. Scale bars: 100µm. (**C**) UMAP of NCAM1-High sorted cells at D70. (**D-E**) CCA integration of scRNA-seq data from embryonic W7.5 eye (D) and NCAM1-High-sorted cells subjected to the altered protocol (E). (**F**) Enriched gene expression for cell types in (C). (**G-H**) Gene expression heatmaps of lens (G) and epithelial (H) cells identified in the reference and *in vitro*. Shared and differentially expressed genes are shown on the left and right plots, respectively. (**I**) RNA velocity of embryonic retinal ganglion cells (left) and hESC-derived neurons (right). (**J-L**) Gene expression analysis of embryonic and hESC-derived neurons along their respective pseudotimes. RGC (Retinal Ganglion Cell), PC (Photoreceptor Cell), HC (Horizontal Cell).

2.3.8. Late differentiation is characterized by the selection and maturation of RPE populations

We next analyzed scRNA-seq at three subsequent time points after D30 replating (D38, D45, and D60) across our three cell lines. Unlike the initial stages, most of the annotated clusters consisted of RPE states. Indeed, the proportion of Late RPE cells was 17.1% ± 14.1 at D38 (5,305 cells), 55.7% ± 13.1 at D45 (5,138 cells), and 77.7% ± 0.6 at D60 (5,777 cells). At the end point of the protocol (D60), approximately 98.2% of cells were at some state of RPE identity, with the remaining fraction consisting of retinal progenitors. Early RPE was characterized by expression of *MITF*, *TYRP1*, *PMEL*, *DCT* while the Late RPE also expressed pigmentation and visual cycle genes *TYR, RLBP1*, *BEST1*, *RPE65*, *SFRP5*, *RGR*, *TTR*, and *RDH5.* At the intermediate time points (D38 and D45), small fractions of non-retinal types, including mesenchyme *(VTCN1, GABRP, HAND1, GATA3)* and smooth muscle contaminants *(MYOG, MYOD1, MYL1, CDH15)* were detected. However, the populations were no longer present in culture at D60 (**Fig. 2.6A-C; Table 2.5**).

Additionally, we also detected a distinct cluster of lingering pluripotent cells in the HS980 D38 sample (Plurip., 0.9% ± 1.1 cells) expressing pluripotency markers *(SOX2, LIN28A, SALL4, GPC3)* (**Fig. 2.6A**). This was concerning as lingering pluripotent stem cells must be eliminated from the final cell product. We therefore extended our analysis including eight independent D60 samples containing 63,370 cells across all three cell lines. Encouragingly, no cells with a pluripotent signature were detected in any of the final D60 samples (**Fig. S2.5A-B, S2.1H-I**).

Interestingly, the D38 time point after replating displayed an increased heterogeneity and on average showed a remarkably less distinct RPE cobblestone morphology than D30 (**Fig. 2.1C, S2.1A**). Furthermore, from D30 onwards a population of RPE cells co-expressed *MITF* and markers associated with the epithelial-to-mesenchymal (EMT) transition process, particularly *ACTA2*. Recent studies have suggested that TGFβ signaling used in RPE differentiation can induce EMT (Boles et al., 2020; Jung et al., 2020; Salero et al., 2012). Despite the absence of Activin A (a TGFβ-superfamily ligand) in culture from D30 and onwards, the dissociation of RPE cells at D30 induced a mesenchymal-like morphology of the RPE cells (**Fig. 2.1B**). This observation led to characterization of two early RPE clusters at most time points from D30 onwards, one *MITF*+*ACTA2*- (EarlyRPE) and one *MITF*+*ACTA2*+ (EMT-RPE). The proportion of EMT-RPE increased during replating from D30 to D38, followed by a steady decrease to low levels (0.8% cells) on D60 (HS980). These cells displayed a signature of some, but not all, RPE markers co-expressed with EMT markers. Moreover, the

representation of RPE from later time points along a phenotype variation axis confirmed the presence of some shared EMT and RPE differentiation properties (**Fig. S2.5C-F**).

Nonetheless, from D30 to D60, we observed the persistence of RPE and loss of other cell types; there was also some maturation variability among the RPE clusters on D60. In fact, pseudotime trajectory inference and RNA velocity analysis of HS980 cells showed a (unidirectional) trajectory of less mature populations in gene expression space towards the most mature RPE (**Fig. 2.6D**). Phase portrait analysis comparing the steady state expectations for spliced and unspliced RNA levels confirmed the upregulation of *RPE65* and *BEST1* as well as the downregulation of progenitor marker *PAX6* (**see Methods; Fig. 2.6E**).

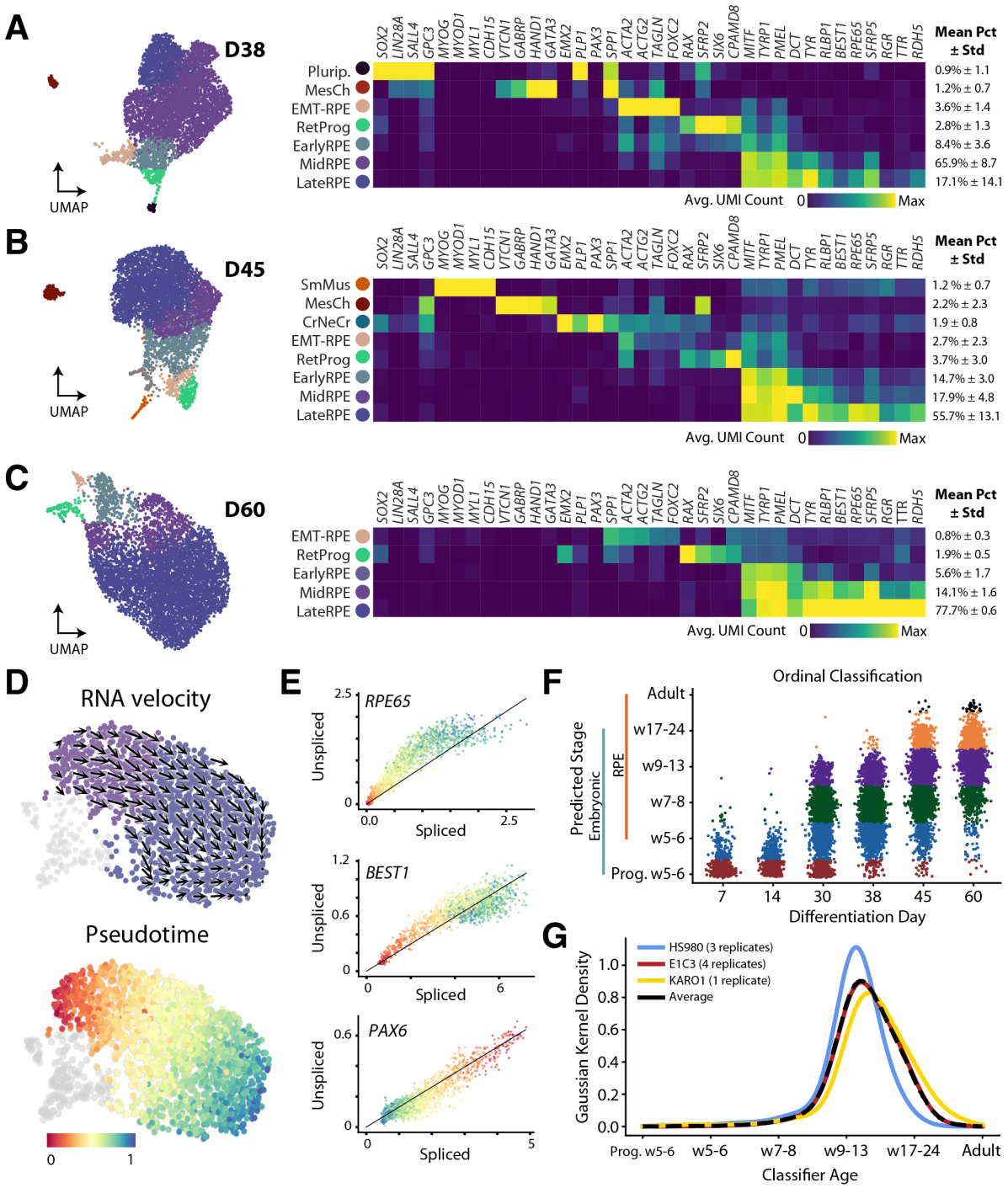


Figure 2.6. Profiling late hESC-RPE differentiation. (**A-C**) UMAP and enriched gene expression heatmap of hESC-RPE scRNA-seq data at D38 (A), D45 (B), and D60 (C) in all three lines. (**D**) RNA velocity and pseudotime analysis of HS980 RPE at D60. (**E**) Phase portraits of upregulated RPE marker genes *RPE65* and *BEST1* as well as a downregulated progenitor marker *PAX6*. The diagonal line represents the estimated steady state of gene expression, with cells above the steady state experiencing gene upregulation and those below gene downregulation. (**F**) Ordinal classification of 20,682 single hESC-derived retinal progenitor and RPE cells at six differentiation time points along embryonic stages. (**G**) Classification distribution for seven hESC-RPE differentiation D60 biological replicates (HS980: 3,655 cells; E1C3: 61,479 cells; KARO1: 1,236 cells). See also Figure S2.6.

2.3.9. Replating affects cell population distribution and promotes a purer and more mature cell product

We have previously shown that replating D30 monolayer cultures greatly facilitates the expansion of final cell numbers (A. Plaza Reyes et al., 2020), but we had not explored how replating affects maturation and purity of the final cell product at D60. We therefore repeated our hESC-RPE differentiation protocol without the replating step and performed scRNA-seq on D38 (1,423 cells) and D60 (793 cells). These data of D38 and D60 non-replated cultures revealed extensive contamination with alternative lineages containing cell types resembling neural retina, neuronal, lens, mesenchyme, neural crest and mesoendodermal cells (**Fig. S2.5G-H**). The presence of contaminant types was confirmed by flow cytometry for lack of the RPE marker CD140b (**Fig. S2.5I**). Interestingly, the absence of replating at D60 also maintained a larger proportion of RetProg (7.2% versus 1.9% in the replated), and there was a lower fraction of LateRPE (33.4% versus 77.7% in the replated) (**Fig. S2.5H, 2.6C**). We determined that replating selects against retinal progenitors and arrests the expansion of alternative lineages. We reasoned that the dissociation and its selection effect might contribute to the formation of a better organized cell monolayer at D60 which, in turn, could promote the acquisition of a more mature RPE expression pattern. This was confirmed by RPE cobblestone morphology quantifications showing higher cobblestone junction scores at D60 replated cultures compared to the non-replated counterparts (**Fig. S2.5J**). Moreover, to assess overall progression of hESC-RPE and assign cells to developmental stages, we constructed an ordinal classifier using single-cell human transcriptomes of 783 embryonic eye cells at W5 to W24 and 127 adult RPE cells (Hu et al., 2019; Voigt et al., 2019) (**Fig. S2.6A-B**). Our classifier operates with an underlying knowledge of the sequential relationship among the training data and was applied to place hESC-RPE cells on the temporal spectrum of RPE development. As proof of principle, we applied our classifier to the human embryonic references at W5 (CS13) and W7.5 (CS20) as well as to an independent set of 49 adult RPEs not used to generate the classifier, thus confirming an appropriate assignment (S. R. Quake & Sapiens Consortium, 2021) (**Fig. S2.6A-C**). Evaluation of the maturity level using such classifier confirmed that replating leads to a more mature output (**Fig. S2.5K**) (Hu et al., 2019). Ultimately, this evidence suggests that replating drives the convergence to a RPE monolayer.

We then decided to analyze the maturation status of all *in vitro* retinal progenitor and RPE cells with our built classifier, and we observed a gradual progression of maturity during differentiation corresponding to embryonic RPE development, which was consistent in all three studied cell lines (**Fig. 2.6F, S2.6D**). Furthermore, the overall classified maturity of D60 cells was similar among the three lines (**Fig. 2.6G**).

Lastly, we sought to compare the classification of RPE cells from our 2D monolayer protocol to (1) D60 RPE cells generated through 3D EB differentiation, and (2) cells from other protocols with longer differentiation. D60 RPE cells in either 3D or 2D differentiation showed similar maturation statuses (**Fig. S2.6H**). In addition, we re-analyzed scRNA-seq datasets in which differentiation was performed for 95 days (9,456 cells) and extended up to 432 days (3,216 cells) using another 2D monolayer protocol originating in the H9 cell line (Lidgerwood et al., 2021) (**Fig. S2.6E-G**). This analysis showed that our D60 cells are similar in maturation status to the D95 cells, but that further maturation can be achieved through extensive *in vitro* culturing to D432 (**Fig. S2.6H**). Interestingly, although both 95- and 432-day time points contain highly mature RPE, these samples also include fractions of retinal progenitors (expressing *CRABP1, SFRP2, PAX6)* and EMT-RPE (expressing *ACTA2, TAGLN)*, which are similar to the progenitors detected in our protocol (**Fig. S2.6G-H**).

2.3.10. Subretinal transplantation of hESC-RPE facilitates a more advanced RPE state

Robust hESC-RPE differentiation is a required first step towards cellular therapies for AMD. However, derived RPE cells must integrate with neighboring tissues upon injection, retain or develop mature attributes, and avoid the resurgence of pluripotent properties to become an effective treatment modality. To evaluate these aspects, D60 hESC-RPE cells were transplanted into the subretinal space of two albino rabbits, a preclinical large-eyed animal model (Bartuma et al., 2015; Petrus-Reurer et al., 2017). Transcriptional analysis of adult rabbit (1,965 cells) and adult human retina (5,538 cells) showed a high degree of similarity (**Fig. S2.7; Table 2.6**). Four weeks following transplantation of hESC-RPE, infrared and SD-OCT imaging showed a pigmented patch of human cells and a hyper-reflective RPE layer among the albino rabbit retinal layers (**Fig. 2.7A**). Histology and immunofluorescence staining further demonstrated that out of all NuMA positive cells (n=227), 99.56% were either pigmented or expressed the RPE marker BEST1, corroborating the successful integration of injected hESC-RPE cells in a polarized and matured RPE monolayer (**Fig. 2.7B, S2.6I-J**). The contiguous injected retina of two rabbits was then processed for scRNA-seq. This yielded 65 human hESC-derived cell profiles, all of which exhibited high expression of mature RPE markers. Crucially, markers of retinal progenitors, photoreceptors, pluripotent hESCs, and EMT-RPE were benchmarked against our references and found to be undetected following transplantation (**Fig. 2.7C**). These findings attest that integrated hESC-RPEs possess the transcriptional signature of mature RPE without signs of retinal progenitor or pluripotent properties.

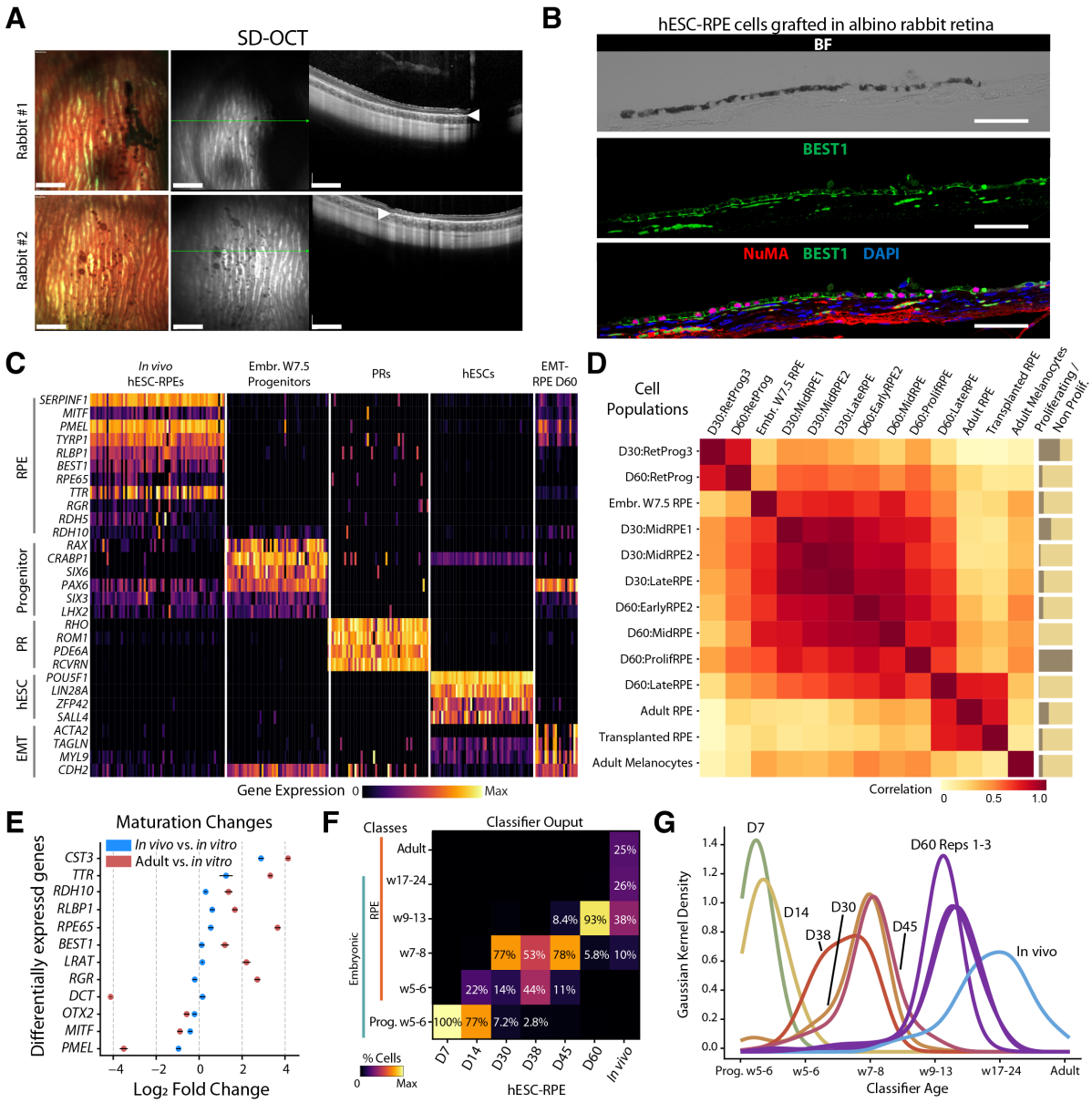


Figure 2.7. Phenotyping of transplanted hESC-RPE. (**A**) Infrared and SD-OCT images of injected hESC-RPE cells in the subretinal space of albino rabbits. Green lines indicate the SD-OCT scan plane. White arrows indicate the hyper-reflective RPE layer. Scale bars: 1mm. (**B**) Brightfield and immunofluorescent staining for human marker NuMA and BEST1 30 days after injection. Scale bars: 50µm. (**C**) Gene expression heatmap comparing 65 single hESC-RPE cells 30 days after transplantation to embryonic W7.5 retinal progenitors, adult photoreceptors, undifferentiated hESCs, and D60 EMT-RPE. (**D**) Pearson's correlation matrix between gene expression profiles of HS980 hESC-RPEs at D30 and D60, post-transplantation (*in vivo*) RPE, adult RPE and melanocytes, and embryonic RPE. (**E**) Dot plot showing log2 fold change of RPE markers between HS980 hESC-RPE D60 cells, *in vivo* RPE, and adult RPE. Bars represent mean +/-SEM from all cells at each time point. (**F**) Ordinal classification summary matrix showing the percentage of HS980 retinal cells from *in vitro* and *in vivo* time points predicted to correspond to each RPE developmental time point (embryonic weeks 5-24, adult). (**G**) Classification distribution for hESC-derived progenitor and RPE cells *in vitro* and *in vivo*. See also Figure S2.7.

We further compared *in vivo* and *in vitro* expression patterns by performing gene expression correlation analysis using both RPE and RetProg clusters from D30 and D60 as well as embryonic reference tissues (**cf. Fig. 2.3, S2.3, S2.6**). Most RPE clusters were well correlated, yet only the most mature *in vitro* cluster (D60:LateRPE) and the *in vivo* transplanted RPE were highly similar to the adult RPE reference. Melanocytes, a distinct neural-crest derived cell type with some overlapping gene expression to RPEs, were not as correlated with hESC-RPE. Interestingly, while D60:LateRPE retained similarities with other *in vitro* clusters, the transplanted RPE did not (**Fig. 2.7D**). Differential expression analysis confirmed an expression pattern closer to adult RPE cells after *in vivo* implantation, particularly for visual cycle components such as *TTR*, *RDH10*, *RLBP1*, *RPE65, BEST1*, and *RGR* (**Fig. 2.7E**).

Lastly, to assess overall progression of hESC-RPE and assign cells to developmental stages, we applied our ordinal classifier to our *in vitro* cells and 65 post-transplantation cells. This showed a gradual progression of maturity during the differentiation time course corresponding to embryonic development that continued *in vivo*. Transplanted hESC-RPE were assigned to late embryonic week 17-24 (26%) and adult RPE classes (25%) more than any *in vitro* time point. Grafted cells were evaluated as more mature than all three HS980 D60 biological replicates, which were predominantly assigned to the week 9-13 stage (93%) (**Fig. 2.7F-G**). Overall, despite the pluripotent state of the cell source and the initial diversity expansion observed, the hESC-RPE differentiation protocol ultimately yielded homogenous and mature RPE cells in a sequence similar to that of embryonic RPE development, and subretinal transplantation of D60 RPE cells assisted with further progression towards a more mature RPE state.

2.4. Discussion

In the present study, by molecularly profiling a directed and defined hESC-RPE differentiation protocol established for clinical translation (A. Plaza Reyes et al., 2020), we demonstrated that the described culture conditions successfully induced RPE lineage specification, selection and maturation over 60 days. Overall, we observe a sequence of gene expression waves consistent with embryological studies (Sabine Fuhrmann et al., 2014; Hu et al., 2019). However, at early stages we found a cell pool heterogeneity that was incompatible with the induction of a single lineage and instead exhibited an initial expansion of cellular diversity (**Fig. 2.1, 2.2**). Similar cell type heterogeneity expansion was previously observed in studies of endoderm and endothelial tissue derivation, but meta-analyses of several differentiation protocols is needed to understand if the observed event is a widespread phenomenon (Cuomo et al., 2020; MacLean et al., 2018; McCracken et al., 2020). Taken

together, characterization of our *in vitro* hESC-RPE differentiation suggests a divergence-convergence model: a diversity expansion at early time points, biased by cell line intrinsic tendency, followed by selection of RPE lineage, driven by replating at D30, and convergence onto a homogeneous and highly pure cellular product.

In the analyzed differentiation protocol, each line manifested different biases to this initial diversification: the E1C3 cells developed endoderm-like cells, HS980 cells produced populations reminiscent of different rostral neural tissues, and a fraction of KARO1 cells displayed signatures of lingering pluripotent cells at the earliest time points (**Fig. 2.2D**). Particularly interesting is the finding of expression profiles resembling patterned regions surrounding the optic field: the pre-placodal epithelium, neural fold, and neural crest (**Fig. 2.2E, 2.2G, S2.2A**). This axis of patterning is induced in the embryo by the organizer cells of the floor plate and anterior neural ridge, which promote specification of different territories of the anterior neural plate, including the optic vesicle (Begbie, 2013a; Eagleson et al., 1995; Sabine Fuhrmann, 2010; Streit, 2007).

These findings hint at an intriguing self-organization process occurring in our 2D culture, despite the lack of spatially directed cues or 3D structure. The investigation of the nature and source of the cues contributing to differentiation phenomenology is a compelling future research direction. For example, it might be interesting to evaluate the potential involvement of neural crest-like and mesenchymal cells appearing during the protocol in the differentiation process, given their known endogenous roles *in vivo* (S. Fuhrmann et al., 2000; Kagiyama et al., 2005).

The heterogeneity at intermediate steps highlights the importance of having robust analysis methods to ensure the final cell product does not contain unwanted cells, such as lingering pluripotent cells that could lead to tumor formation. In this respect, the consistent output obtained across different cell lines and the achievement of a mature endpoint, even when rebooting the protocol at D30 from NCAM1-High sorted populations, constitute an important proof that the present differentiation method efficiently eliminates such impurities. The comparison of our protocol to EB 3D differentiation further contextualizes the extent of convergence induction in a broader setting. While the 3D setting promotes a wider patterning and the simultaneous development of different neural and retinal lineages, where RPE cells are only one of the outputs, our 2D culture conditions successfully facilitate a complete convergence on mature RPE by D60. Furthermore, the disappearance in later time points of the small smooth muscle and myogenic cell populations temporarily emerging at D38 hints at how the combination of a selective pressure component, such as the replating step, could contribute to this convergence (**Fig. 2.6, S2.5**).

In this work, we used human embryonic and adult references to evaluate the composition and level of maturity of the cells we generate, and to objectively compare our results to the output of other published protocols. The results highlight that an adult RPE pattern of expression is not yet reached in cells at D60, and a more mature pattern can be achieved at the cost of one year of further culturing (Lidgerwood et al., 2021). Intriguingly, even these long-term matured RPE cultures revealed, at reanalysis, a comparable fraction of retinal progenitors and an EMT-RPE persisting in culture. Further studies are warranted to elucidate the function of these populations to understand if they may represent a normal part of RPE physiology and how they may impact cell therapy products.

Nonetheless, efficient sorting procedures for cell population purification at intermediate time points could lead to the design of differentiation protocols reaching full maturity faster. In this respect, we showed that removing the NCAM1-High progenitor pool at D30 yields a more mature RPE population in a shorter period of time (**Fig. 2.4, S2.4**). At the same time, our data also show that the replating itself reduces the fraction of retinal progenitors in the final product as well as other contaminating cell types, reaching a purity of 98% at D60, where the remaining non-RPE is a 2% retinal progenitor fraction. Such purity is well in line with or greater than other hESC-based cellular therapies (Piao et al., 2021). Furthermore, considering the added manufacturing challenges with an antibody-based sorting, such enrichment would not be cost-effective in this setting. In the present work, we also demonstrated that NCAM1-High cells are not RPE-fate restricted and, upon altered culture conditions, can give rise to different cell types, including anterior neurons, mesenchyme, and lens epithelium (**Fig. 2.5**). This potency is particularly relevant, as the identification and isolation of less mature progenitors with an increased plasticity is of importance to efforts aimed at replacement of other retinal cell types affected by advanced AMD (Bhatia et al., 2010; Marquardt et al., 2001). Thus, further evaluation of the NCAM1-High potency as a response to different and more specific culture conditions, and in other *in vivo* models lacking certain retinal cell types constitute promising avenues for future investigation.

Considering that the final cell product may contain 2% of such retinal progenitors, it should be averted in the future despite being unlikely to pose a safety risk. However, we did not detect any non-RPE cell types from our histological or transcriptional analysis following cell transplantation (**Fig. 2.7, S2.6I-J**). Although the number of cells analyzed (227 by tissue immunofluorescence and 65 transcriptionally) is admittedly not very large, these findings suggest that the progenitor pool has not expanded extensively or generated alternative lineages. Additionally, our in-depth analysis highlights the importance of ensuring that the final cell product does not contain lingering pluripotent stem cells at a single-cell level. We clearly

saw initial numbers of remaining pluripotent-like cells in D7 and D14 cultures, but these were largely lost at D30. With this basis, it was surprising to detect a distinct population re-emerging at D38. Importantly, our focused analysis showed that at D60, none of the eight samples from the three cell lines (63,370 cells) contained cells with a transcriptional profile corresponding to pluripotency (**Fig. 2.6, S2.6**).

The behavior of grafted cells *in vivo* is a topic discussed extensively by the community, with maintenance of the proliferative potential and de-differentiation generally considered the two processes of major concern (Wang et al., 2020; Zarbin et al., 2019). Our analysis identified neither specific signs of de-differentiation nor the presence of a non-RPE molecular profile. Indeed, we detected a distinct shift in the RPE maturation towards a more adult and functional phenotype (**Fig. 2.7**). The induction mechanism of the observed *in vivo* maturation remains unclear; albeit, the increased expression of visual cycle genes suggests that donor cells support neighboring photoreceptors functionally.

Overall, our findings provide a high-resolution perspective on human pluripotent stem cell differentiation and a comprehensive and necessary detailed analysis of a stem cell-based product intended for successful and safe human therapeutic strategies. Ultimately, this study will guide future efforts focused on the differentiation of retinal cells, a deeper understanding of mechanisms of retinal disease, and applications in regenerative medicine.

2.5. Methods

For the complete extended methods to the findings presented in this chapter, please refer to the original publication (Petrus Reurer, Lederer, et al., 2022) in *Stem Cell Reports* at: https://doi.org/10.1016/j.stemcr.2022.05.005.

2.5.1. hESC Cell Culture and hESC-RPE Differentiation

hESC lines HS980, KARO1 were established and cultured in 5% CO_2/5% O_2 on rhLN-521 (10µg/mL), and passaged as described previously (Rodin et al., 2014). E1C3 (NN GMP0050E1C3) cultured on iMatrix-511 (0.25 µg/cm^2, Nippi, T303) was provided as a research cell bank of the clinical GMP cell line by NovoNordisk (UCSF IRB: 1518222, for RPE differentiation Projekt-ID: H-18016740, Anmeldelsesnr.: 73105).

For differentiation (Plaza Reyes et al. 2020a and 2020b), cells were plated at a density of 2.4x10^4 cells/cm^2 on 20µg/mL hrLN-521 or iMatrix-coated dishes using NutriStem hPSC XF medium and Rho-kinase inhibitor (10uM) during the first 24h. Medium was then replaced with NutriStem hPSC XF without bFGF and TGFβ (differentiation medium) in 5% CO_2/21% O_2, and from day 6 after plating, 100 ng/mL of Activin A was added to the medium for a total of 30

days. D30 monolayers were replated using TrypLE Select (10 min, 37°C) and passed through a 40um strainer. Cells were seeded on hrLN-521 coated dishes (20µg/mL) at $6.8x10^4$ cells/cm^2, and fed three times a week for subsequent 30 days with differentiation medium without Activin A.

2.5.2. Sample Processing for Single-Cell RNA Sequencing

Cells: Specific stage hESC-RPE cells were trypsinized with TrypLE (10 min, 37°C, 5% CO_2) and resuspended to 1000 cells/µL in 0.04% BSA in PBS prior to scRNA-seq. Tissues: Two human 32h post-mortem eyes from the same donor were collected, retinas were dissected out and cut into several small pieces mixed together in 500 µL of digestion buffer (see Supplemental Experimental Procedures). Two pooled embryonic eyes at Carnegie Stages 12, 13, 14, 15 (5 post-conception week), and two embryonic eyes from the same donor (7.5 post-conception week) were collected. Optic cups were dissected out and chopped in several small pieces to facilitate dissociation in 500 µL of digestion buffer. Two rabbit eyes (from different animals) with 30-day integrated hESC-RPE were enucleated and neuroretina, choroid and RPE layer, were dissected out and mixed together in 500 µL of digestion buffer. After digestion (37°C, 25 min on a 300g rotator, resuspended every 5 min), samples were filtered using a 30µm strainer followed by Dead Cell Removal kit. At this stage, one of the rabbit eye cell samples was stained with mouse anti-human HLA-ABC-FITC; HLA-ABC-positive cells were FACS-sorted, collected and resuspended to 1000 cells/µL in 1% BSA in PBS. The rest of the samples were also resuspended to 1000 cells/µL in 1% BSA in PBS prior to scRNA-seq.

2.5.3. Single-Cell RNA Sequencing Analysis

Cells were either transported at 4°C to the Eukaryotic Single Cell Genomics Facility (ESCG, SciLifeLab, Stockholm, Sweden) or used in-house to prepare cDNA libraries for scRNA-seq. The 10X Genomics Single Cell 3' Reagent Dual Index Kit v2 and v3.1 (10x Genomics, CG000315) was used, sometimes with an additional Cell Multiplexing Oligo Labeling step (10x Genomics, CG000391), followed by protocol CB000388 and sequencing on a NovaSeq 6000 (ESCG) or Illumina Nextseq 2000 (in-house). Cell Ranger 3.1.0 was used to convert base call files to FASTQ format, map sequencing reads to the human GRCh38 reference transcriptome, and generate feature-barcode matrices. For the E1C3 cell line sequenced at NovoNorDisk, CellRanger 3.0.2 was used. Quality control, normalization, dimensionality reduction, and visualization were performed using the *scanpy* and *velocyto* modules (La Manno et al., 2018; Wolf et al., 2018). For samples on which RNA velocity was

performed, the *velocyto run10x* command was used on CellRanger sorted BAM files to produce loom files containing spliced and unspliced counts. Cell filtering, dimensionality reduction and visualization criteria are provided for each individual sample in the Supplemental Experimental Procedures and **Table 2.1**.

2.5.4. Subretinal Transplantation and In Vivo Imaging

Dissociated hESC-RPEs were injected (50 µL; 50,000 cells) subretinally using a transvitreal pars plana technique in New Zealand white albino rabbits (Bartuma et al., 2015; Petrus-Reurer et al., 2017, 2018). SD-OCT and confocal scanning laser ophthalmoscopy was performed to obtain horizontal cross-sectional B-scans and en face fundus in vivo images, respectively.

2.5.5. Data and Code Availability

FASTQ files, processed feature-barcode count matrices, annotated h5ad/loom files, and other metadata are available on GEO (GSE164092). Jupyter notebooks for the single-cell analyses are shared at https://github.com/lamanno-epfl/rpe_differentiation_profiling_code. Datasets are available for interactive visualization and analysis at https://asap.epfl.ch/ under public keys ASAP 75-90 (David et al., 2020).

2.5.6. Acknowledgements

We thank Ernest Arenas, Pierre Fabre, Igor Adameyko, Pierre Gönczy, Felix Naef and Bart Deplancke for helpful feedback on the manuscript. We also thank the EPFL Histology Core Facility team, particularly Jessica Dessimoz, Gian-Filippo Mancini, and Nathalie Müller, for their assistance with immunostainings. The work was supported by grants from the Swedish Research Council, Ragnar Söderberg Foundation, Ming Wai Lau Center for Reparative Medicine, Center for Innovative Medicine, Wallenberg Academy Fellow, Strategic Research Area (SRA) Stem Cells and Regenerative Medicine, Vinnova, Stockholm County Council (ALF project), Karolinska Institute, Crown Princess Margareta's Foundation for the Visually Impaired, ARMEC Lindeberg Foundation, the Ulla och Ingemar Dahlberg Foundation, and King Gustav V and Queen Victoria Foundation, Cronqvist Foundation, the Swiss National Science foundation grants CRSK-3_190495 and PZ00P3_193445, and grant CZF2019-002427 from the Chan Zuckerberg Initiative.

This study was performed at the Live Cell Imaging unit/Nikon Center of Excellence, BioNut, KI, supported by Knut and Alice Wallenberg Foundation, Swedish Research Council, Centre for Innovative Medicine and the Jonasson donation. Flow cytometry was performed at

the MedH Flow Cytometry core facility, and prenatal human tissue was acquired through the Developmental Tissue bank core facility, both supported by KI/SLL. Sequencing was performed at ESCG Infrastructure in Stockholm at Science for Life Laboratory (funded by the Knut and Alice Wallenberg Foundation and the Swedish Research Council) and Bioinformatics and Expression Analysis (supported by the board of research at the Karolinska Institute and the research committee at the Karolinska Hospital) with assistance from SNIC/Uppsala Multidisciplinary Center for Advanced Computational Science with massively parallel sequencing and access to the UPPMAX computational infrastructure.

2.5.7. Author Contributions

S.P.-R., A.R.L., F.L., G.L.M. conceived the study; F.L., G.L.M., J.C.V supervised the work. S.P.-R., L.B.-V., I.K., M.W., H.A., E.S., A.B., A.W., Y.S., P.E., A.Kr. and A.Kv. performed experiments; I.D. and B.P. helped with the cell sorting. H.B. and M.A. contributed to the animal work; A.R.L., E.J., H.W. and G.L.M. performed single-cell RNA sequencing analysis; S.P.-R., A.R.L., A.K., G.L.M. and F.L. planned experiments, analyzed data and wrote the manuscript. S.P.-R. and A.R.L. contributed equally to this study.

2.5.8. Declaration of Interests

S.P.-R., and F.L. are the inventors of a patent (Methods and compositions for producing retinal pigment epithelium cells, filed 19.06.2019, PCT/EP2019/066285). All authors declare no other competing interests.

2.6. Supplementary Materials

2.6.1. Supplementary Figures

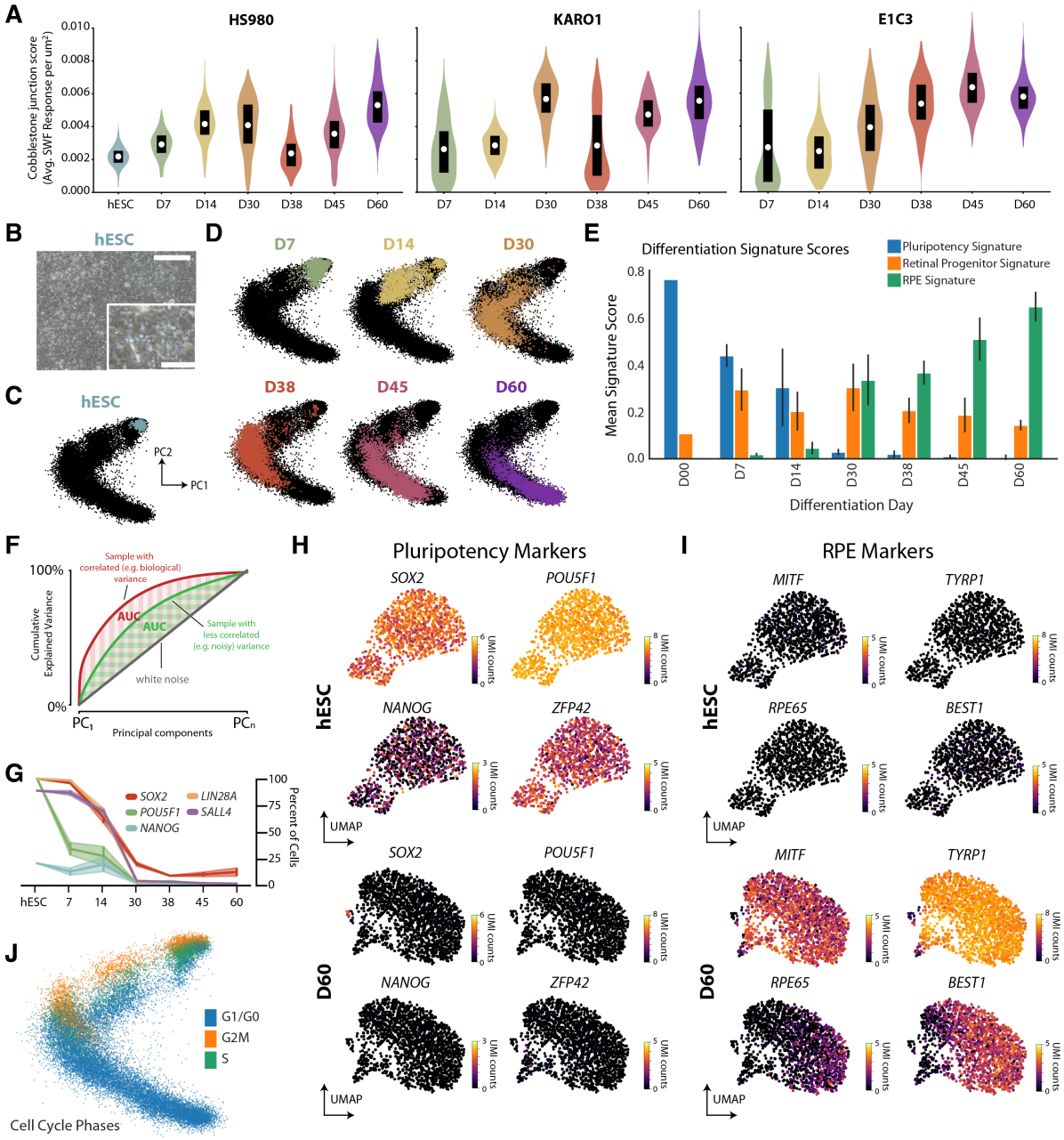


Figure S2.1. Cellular heterogeneity analysis of hESC-RPE differentiation. Related to Figure 2.1. **(A)** Graphs showing quantification of cobblestone morphology throughout differentiation in the HS980, KARO1, and E1C3 cell lines using the junction score methodology and software developed by Joshi et al., 2016. **(B)** Brightfield image of undifferentiated hESCs in the HS980 line. Scale bars: 100μm; inset 20μm. **(C)** Principal component (PC) representation of hESCs in the HS980 line. **(D)** PC representation of *in vitro* hESC-RPE time points across three lines, colored by day. **(E)** Bar graph of average pluripotency, retinal progenitor and RPE signature scores by differentiation day. Error bars represent standard deviation of the mean over three cell line replicates. **(F)** Schematic of AUC variance evaluation metric. **(G)** Graph showing the percentage of cells positive (>0.5 normalized UMI counts) for pluripotency marker genes at each time point. **(H, I)** UMAPs showing normalized gene expression of pluripotent stem cells markers (H) and RPE markers (I) in undifferentiated hESCs and at D60. **(J)**

Principal component representation of hESC-RPE differentiation across all lines colored by assigned cell cycle phase.

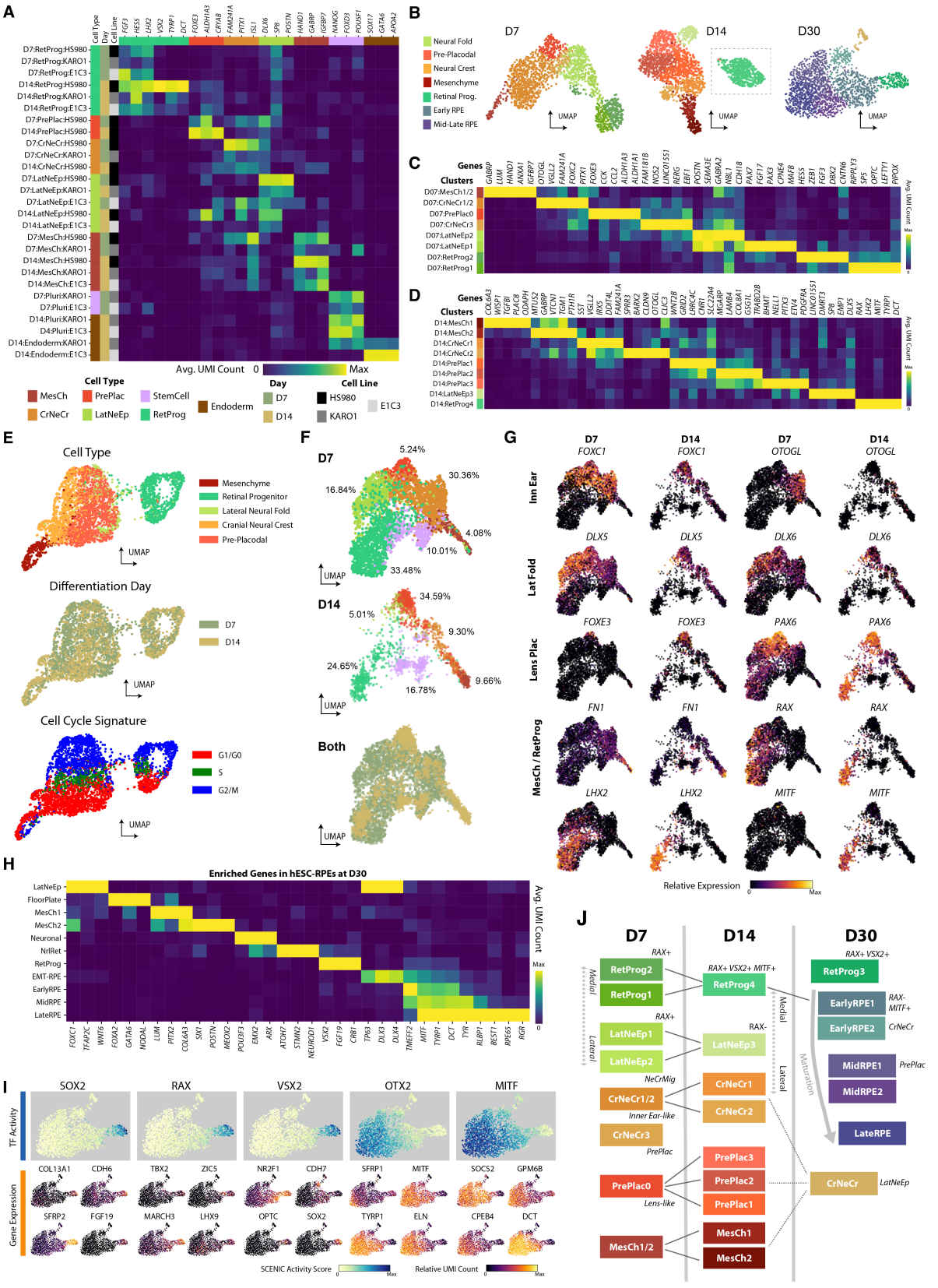


Figure S2.2. Gene expression characterization and canonical correlation analysis of early differentiation. Related to Figure 2.2. (A) Heatmap of enriched genes by primary clusters, grouped by cell line, at D7 and D14 of hESC-RPE differentiation. **(B)** UMAP representation of HS980 differentiation at D7, D14, and D30, colored by cell type. **(C)** Heatmap of top enriched genes of each cell type cluster at D7. **(D)** Heatmap of top enriched genes of each cell type cluster at D14. **(E)** D7 and

D14 cell HS980 populations projected on a shared low dimensional subspace using canonical correlation analysis (CCA; see Experimental Procedures), colored by cell type, differentiation day, and cell cycle phase. **(F)** UMAP representation of D7 and D14 cells, across all three lines, integrated with CCA. **(G)** UMAPs showing gene expression in all three lines of fundamental cell type markers for Inner Ear (InnEar), Lateral Fold (LatFold), Lens Placode (LensPlac), Mesenchyme (MesCh), and Retinal Progenitor (RetProg) from D7 to D14. Expression of neural crest inner ear (*FOXC1, OTOGL*) and lateral fold (*DLX5, DLX6*) markers decreases from D7 to D14. **(H)** Heatmap of enriched genes by cell type at D30 across all three lines. **(I)** Top: Transcription factor (TF) activity scores for SOX2, RAX, VSX2, OTX2, and MITF obtained by SCENIC analysis. Bottom: UMAPs showing gene expression of the top four inferred target genes of each TF at D30 of differentiation in HS980. **(J)** Schematic of the proposed relationship among the various secondary clusters during pigmentation induction. Edges indicate putative relationships between cell types identified at different time points.

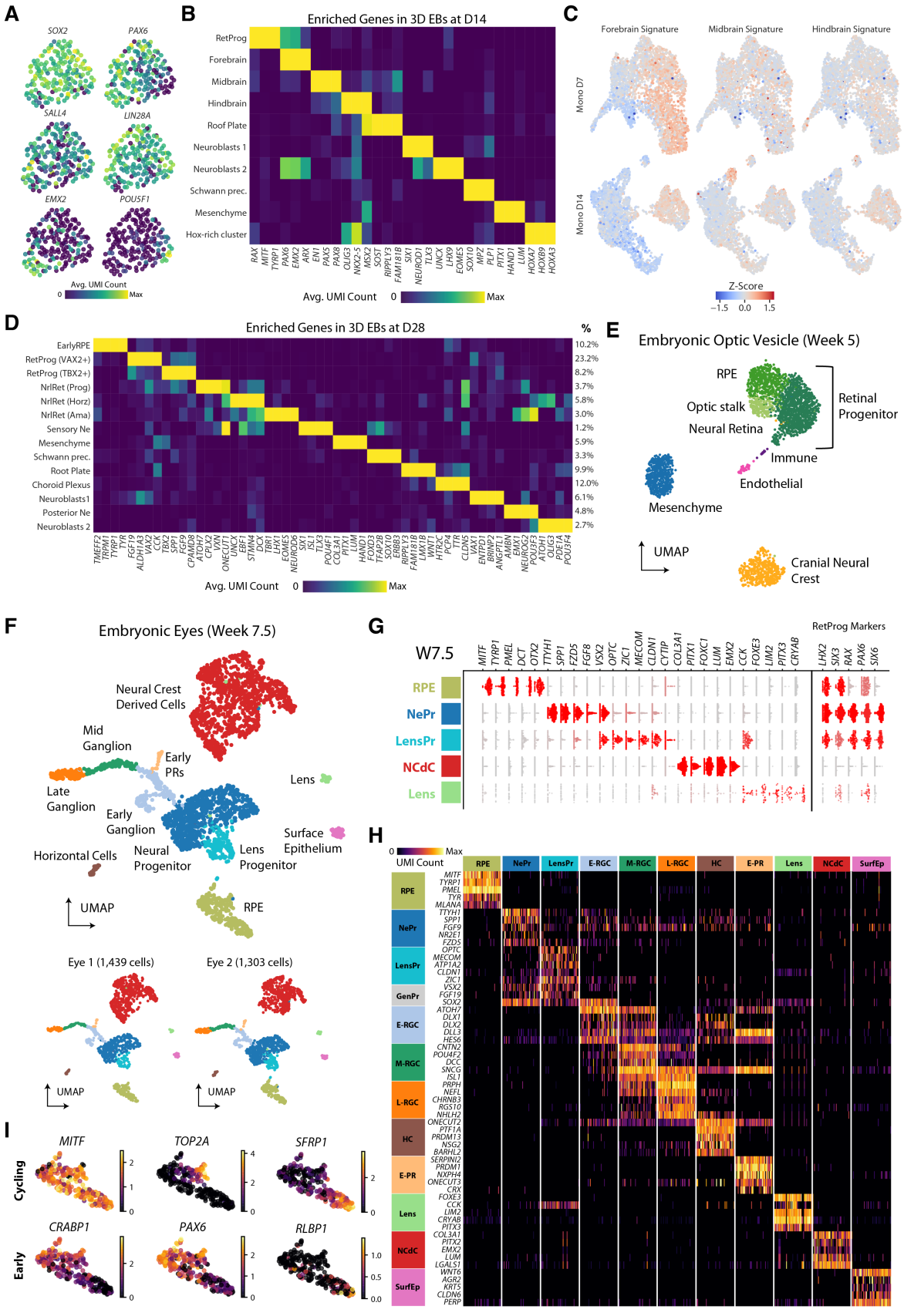


Figure S2.3. Characterization of RPE differentiation in 3D embryoid bodies and compared to embryonic references. Related to Figure 2.3. (A) UMAP representation of EB differentiation a D7 (181 cells), colored by normalized UMI count of progenitor (*SOX2, PAX6*), pluripotency (*SALL4,*

LIN28A) and regional (*EMX2, POU5F1*) marker genes. **(B)** Heatmap of enriched genes by cell type at EB at D14. **(C)** Signature scores for forebrain, midbrain and hindbrain visualized on D7 and D14 monolayer cells. Scores were computed as those in Figure 2.3B. **(D)** Heatmap of enriched genes by cell type at EB at D28, with percent composition of total EB population. **(E)** UMAP representation of a human embryonic optic vesicle (2,637 cells) dissected at 5 weeks (Carnegie Stage 13). Cluster identities include: optic cell types derived from retinal progenitors (RetProg), such as retinal pigment epithelium (RPE), neural retina (NR), and optic stalk (OS), in addition to periocular mesenchyme (MesCh), cranial neural crest (CrNeCr), immune, and smooth muscle. **(F)** Top: UMAP representation of scRNA-seq data from two human fetal eyes at week 7.5, colored by cell type. Bottom: UMAP of each individual fetal eye separately. **(G)** Violin plots of enriched genes in the identified W7.5 clusters. **(H)** Heatmap of normalized enriched gene expression. Genes were selected using an enrichment score by cell type in (F). **(I)** Retinal progenitor and RPE log2 normalized gene expression of cycling, Early and Mid RPE markers in the RPE cell cluster from (F).

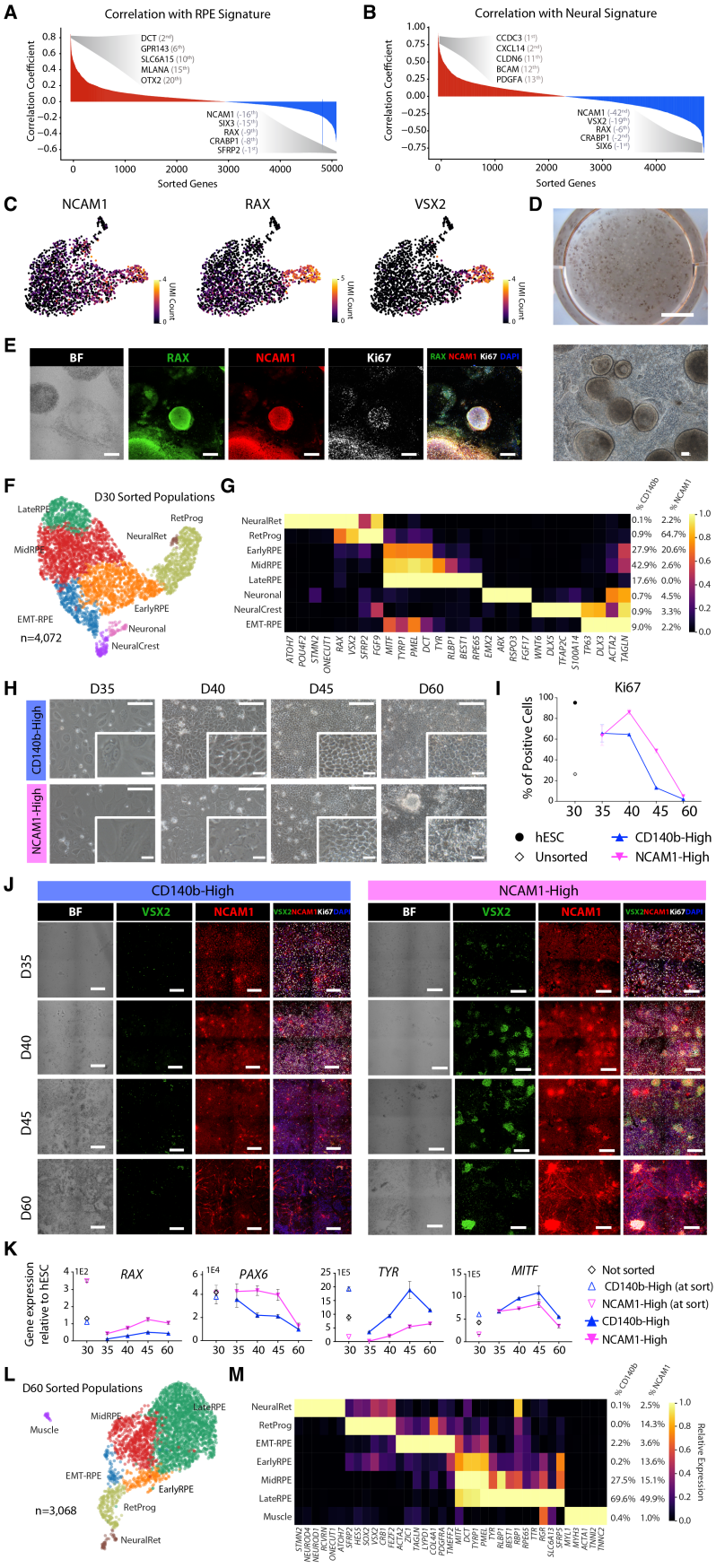


Figure S2.4. Characterization of CD140b-High and NCAM1-High sorted populations exposed to RPE differentiation conditions. Related to Figure 2.4. (A-B). Pearson's correlation coefficients were

computed between 5,412 and an RPE signature (A) or 4,664 genes and a neural signature (B) using normalized counts and cells belonging to the D30 retinal progenitor and RPE clusters or retinal progenitor and neural clusters, respectively. **(C)** UMAPs showing gene expression of progenitor markers in scRNA-seq hESC-RPE at D30 in the HS980 cell line. **(D)** Camera pictures of hESC-RPE D30 cultures. **(E)** Brightfield and immunofluorescence stainings of hESC-RPE D30 cells showing co-expression of RAX, NCAM1 and Ki67 markers. Scale bars: top 1mm; bottom 100μm. **(F)** scRNA-seq of 4,072 single cells from NCAM1-High sorted, CD140b-High sorted, and unsorted D30 cells, colored by cell type. See Figure 4E for composition by sample identifier. **(G)** Heatmap of enriched marker genes in cell types of sorted populations at D30, with cell type composition percentages for both sorted populations. **(H)** Brightfield images of unsorted, CD140b-High and NCAM1-High populations at D33, D40, D45, and D60. FACS sorting of the two populations was performed at D30. Scale bars: 100μm; inset 20μm. **(I)** Graph showing the percentage of positive cells expressing the Ki67 proliferation marker in hESC, unsorted, CD140b-High and NCAM1-High populations at the moment of FACS sorting (D30) and D35, D40, D45, and D60. Bars represent mean +/-SEM from three independent experiments. **(J)** Brightfield and immunofluorescence images showing expression of VSX2 and NCAM1 in unsorted, CD140b-High and NCAM1-High populations after FACS sorting at differentiation D35, D40, D45, and D60. Scale bars: 200μm. **(K)** Graphs representing RT-qPCR of retinal progenitor *(RAX, PAX6)* and RPE *(MITF, TYR)* marker genes in unsorted, CD140b-High and NCAM1-High populations at the moment of sort and at post-sort D30, 35, 40, 45, and 60. **(L)** scRNA-seq of 3,068 single cells from NCAM1-High sorted, CD140b-High sorted, and unsorted D60 cells, colored by cell type. See Figure 4K for composition by sample identifier. **(M)** Heatmap of enriched marker genes in cell types of sorted populations at D60, with cell type composition percentages for both sorted populations.

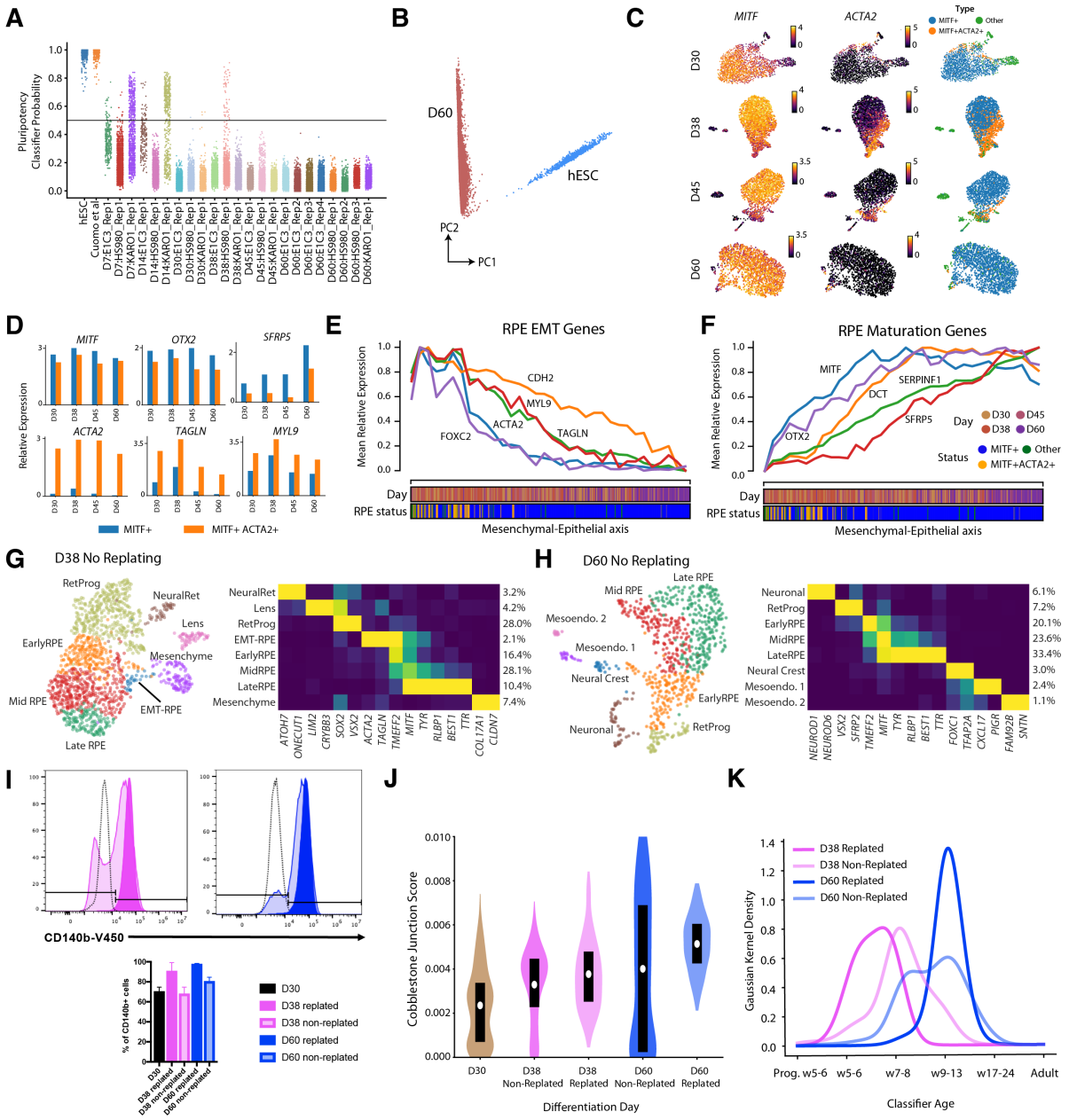


Figure S2.5. Characterization of late differentiation and overall gene expression correlation. Related to Figure 2.6. (A) Plot displaying the probability outputted by a pluripotency classifier when inputted data from in vitro cells at all time points. A random forest classifier was trained on a mixture of hESCs from this and another study (Cuomo et al 2020; see Experimental Procedures) **(B)** Scatter plot of the two first principal components of all D60 and hESC in vitro cells. **(C)** UMAP overlayed with MITF and ACTA2 gene 11 expression at D30, D38, D45, and D60. Cells are labeled as MITF+ (maturing RPE), MITF+ACTA2+ (EMT-RPE), or other (non-RPE cell types). **(D)** Bar graphs showing the gene expression differences between EMT-RPE and maturing RPE. EMT-RPE expresses EMT markers *ACTA2*, *TAGLN*, and *MYL9* more highly, whereas RPE markers *MITF*, *OTX2*, and *SFRP5* are more highly expressed in non-transitioning RPE. **(E-F)** Line plots showing average expression of RPE-EMT (E) and mature RPE (F) along a mesenchymal-epithelial axis of variation determined by fitting a principal curve (see Experimental Procedures). Colored bars on the x-axis indicate time point and RPE status of cells along the axis. **(G)** Left: UMAP representation of 1,423 single cells at hESC-RPE D38 without replating at D30. Right: heatmap of enriched genes by cell type. **(H)** Left: UMAP representation of 772 single cells at hESC-RPE D60 without replating at D30. Right: heatmap of enriched genes by

cell type. **(I)** Top: representative flow cytometry plots for HS980 cell line showing CD140b cell surface marker expression at D38 and D60 for replated and non-replated conditions. Dotted lines represent hESC (negative control). Bottom: bar graphs show the average of CD140b marker expression in the stated conditions and time points for both HS980 and KARO1 cell lines. **(J)** Violin plot displaying the quantification of cobblestone morphology at D38 and D60 with and without replating in the HS980 cell lines using the junction score methodology and software developed by Joshi et al, J Ocul Pharmacol Ther. 2016. **(K)** Graph showing distribution of classifications of D38 and D60 RPE cultures, with and without replating (see also Figures 2.6 and S2.6). Bars represent mean +/-SEM from three independent experiments.

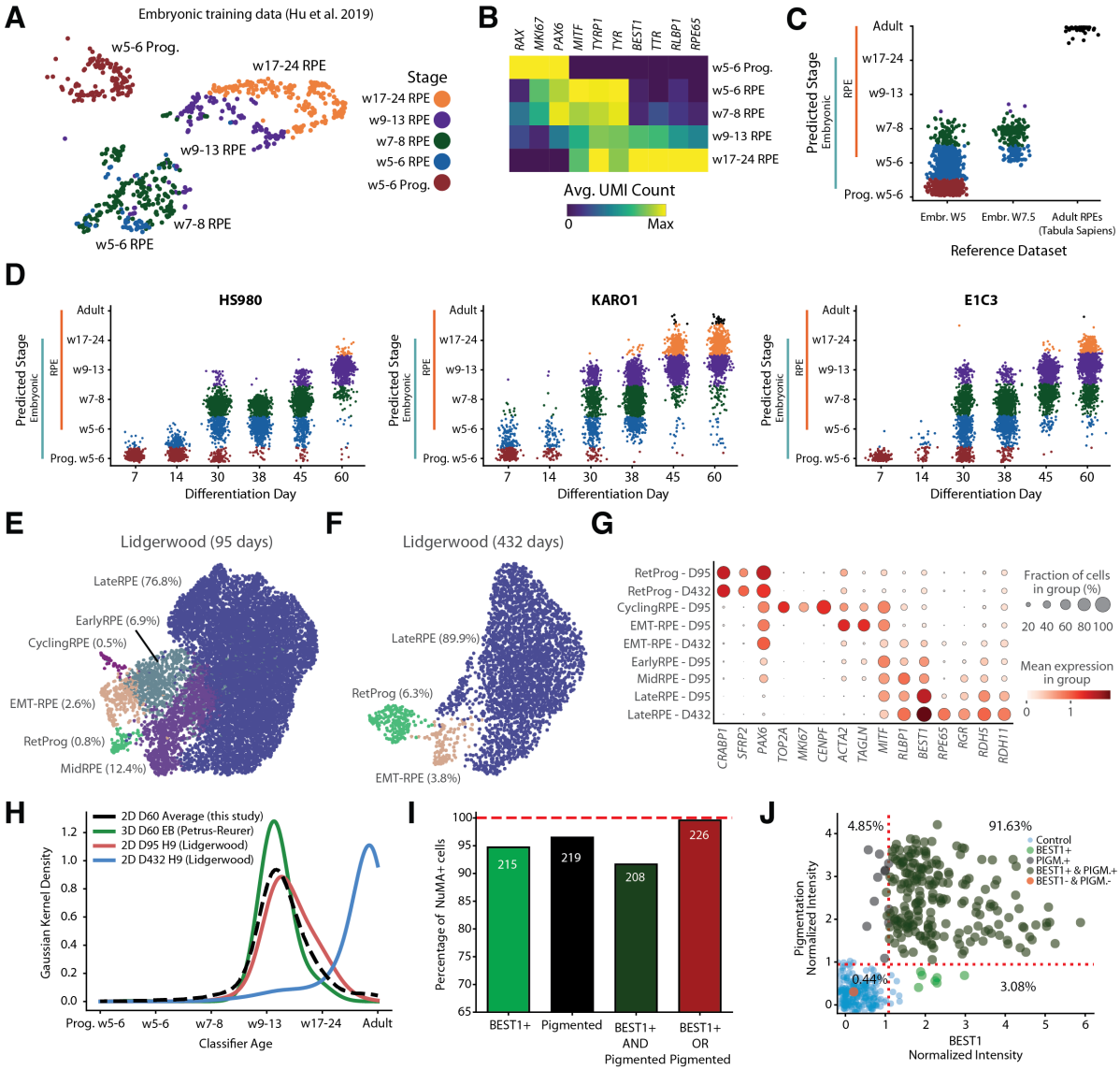


Figure S2.6. Ordinal classification of in vitro hESC-RPE. Related to Figure 2.6. (A) UMAP representation of Hu et al dataset of 783 human fetal cells from various time points in development, as used in the ordinal classifier (see Figure 2.7). Cells were colored into five categories for classification. **(B)** Heatmap showing an overview of uniquely enriched retinal progenitor and RPE marker genes in the data in (D). **(C)** Plot showing ordinal classification of reference embryonic RPEs at weeks 5 and 7.5 (see Figures 2.3 and S2.3). **(D)** Plots showing ordinal classification for hESC-RPE differentiation data in the HS980, KARO1, and E1C3 cell lines individually. **(E)** UMAP representation of RPE differentiation 13 day 90 in H9 cell line (9,456 single cells) re-analyzed from Lidgerwood et al, Genomics Proteomics Bioinformatics 2020. Cells colored and labeled by newly-annotated cell types. **(F)** UMAP representation of RPE differentiation day 432 (1 Year) in H9 cell line (3,216 single cells) re-analyzed from Lidgerwood et al, Genomics Proteomics Bioinformatics 2020. Cells colored and labeled by newly-annotated cell types. **(G)** Dot plot of marker gene expression for RetProg, CyclingRPE, EMT-RPE, EarlyRPE, MidRPE, and LateRPE in (A-B). (H) Graph showing distribution of ordinal classifications of various differentiated RPE culture protocols, including the 2D monolayer protocol from this study (20,682 cells), 3D EBs at D60 (Petrus-Reurer et al, 2020; 294 cells), 2D D95 H9 (Lidgerwood et al., 2021; 9,456 cells) and 2D D432 H9 (Lidgerwood et al., 2021; 3,216 cells). **(I)** Bar graph of 227 hESC-RPE grafted cell BEST1 and pigmentation statuses after 30 days into the albino rabbit subretinal space. NuMA+ human cells from ten sections and three rabbits were manually segmented and assessed by immunofluorescence

(BEST1 expression) and brightfield (Pigmentation). **(J)** Scatter plot of the BEST1 normalized intensity and the pigmentation normalized intensity for 227 grafted NuMA+ cells and 227 NuMA- control cells. Cells are colored by histological status (Control, BEST1+, PIGM+, BEST1 & PIGM+, and BEST1- & PIGM-). Red dotted lines indicate the 97.5th percentile threshold of the signal observed in the negative control cells.

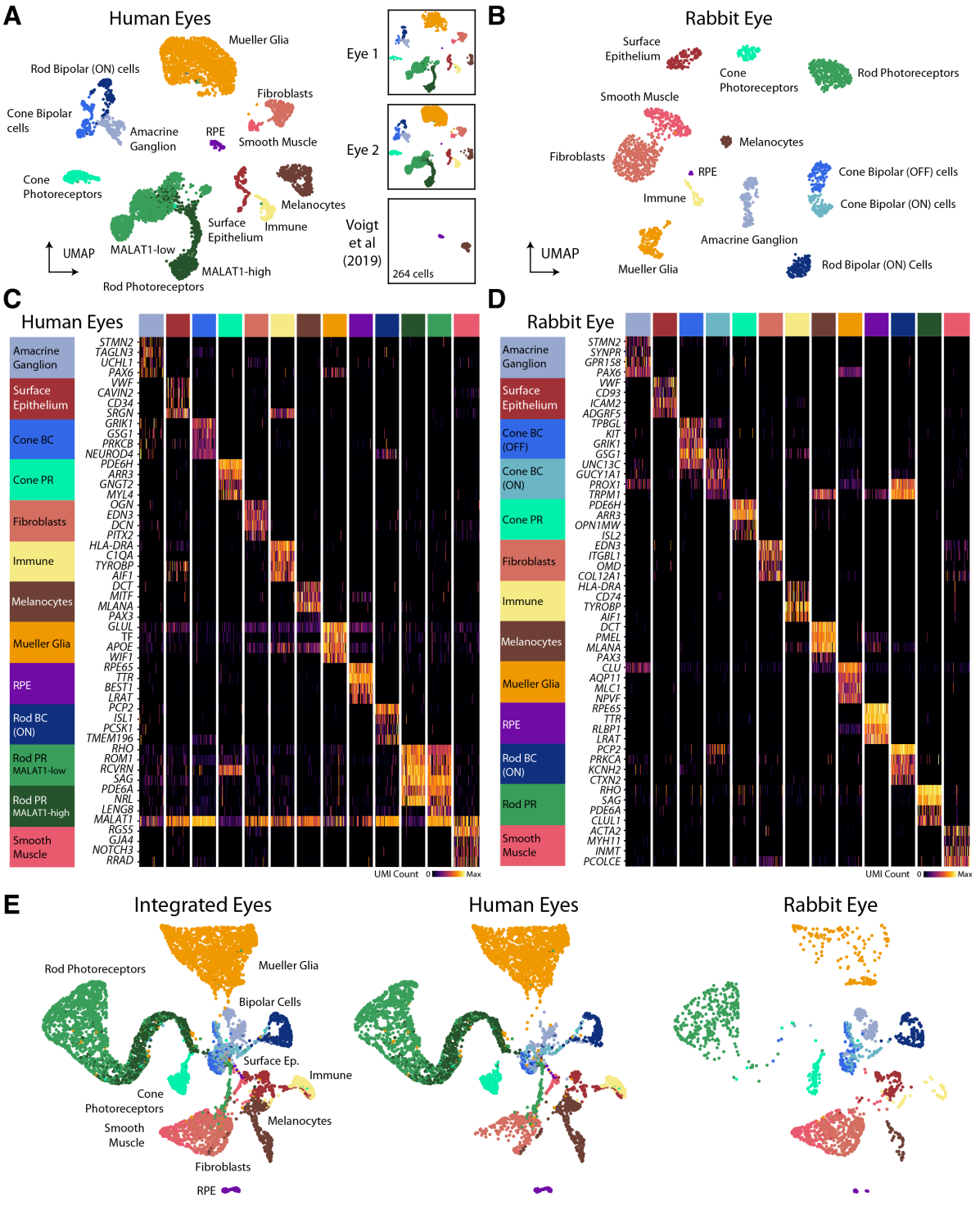


Figure S2.7. Transcriptional analysis of albino rabbit and human retinas. Related to Figure 2.7. **(A)** Annotated UMAP representation of 5,538 human cells from two human 15 eyes (1,564 cells and 3,706 cells), categorized into 13 different cell types. Additional RPEs and melanocytes (264 cells) were re-analyzed and incorporated from Voigt et al, 2019. **(B)** Annotated UMAP representation of 1,965 rabbit cells categorized into 13 different retina cell types. **(C)** Heatmap of enriched marker genes for adult eye cell types. **(D)** Heatmap of enriched marker genes for rabbit eye cell types. Genes were selected from among the top 20 enriched genes per cluster for (C) and (D). **(E)** CCA integration of human and rabbit eyes. Integration was performed using Seurat on 2,000 enriched genes from a total of 9,889 genes with shared annotations between the two species (see Methods 2.5).

2.6.2. Supplementary Tables

Please refer to the published article for the relevant supplemental tables:
https://doi.org/10.1016/j.stemcr.2022.05.005

2.7. Appendix

This section briefly presents some ongoing follow-up work in collaboration with the Lanner Lab, led by Laura Baqué Vidal, to transcriptomically compare hESC-derived RPE cells after cryopreservation. Here, we performed scRNA-seq of RPE that were differentiated for 60 days and re-plated for three days before cryopreservation; this second replated step was found to be necessary for cell viability with cryopreservation. To understand this finding, we compared these D60+3 (D63) cells to D60 cells from the original study to investigate whether the RPE cells are of equal maturity and whether there are any other sub-populations that emerge. From my analyses, it seems that D60+3 (D63) cells tend to de-differentiate slightly, with a larger fractions of EMT-RPE and increased cell cycle signatures (**Fig. A2.1**). No other off-target populations arise. These findings suggest that hESC-derived RPE retain enough plasticity when replated to be suitably cryopreserved. Furthermore, replating may specifically allow for this by facilitating an increase in expression of EMT-related marker genes typical of an earlier RPE state (**Fig. S2.5**). A similar phenomenon was observed in the original study after D30 replating (**Fig. 2.6, S2.6**). For more considerations, please see Chapter 5.

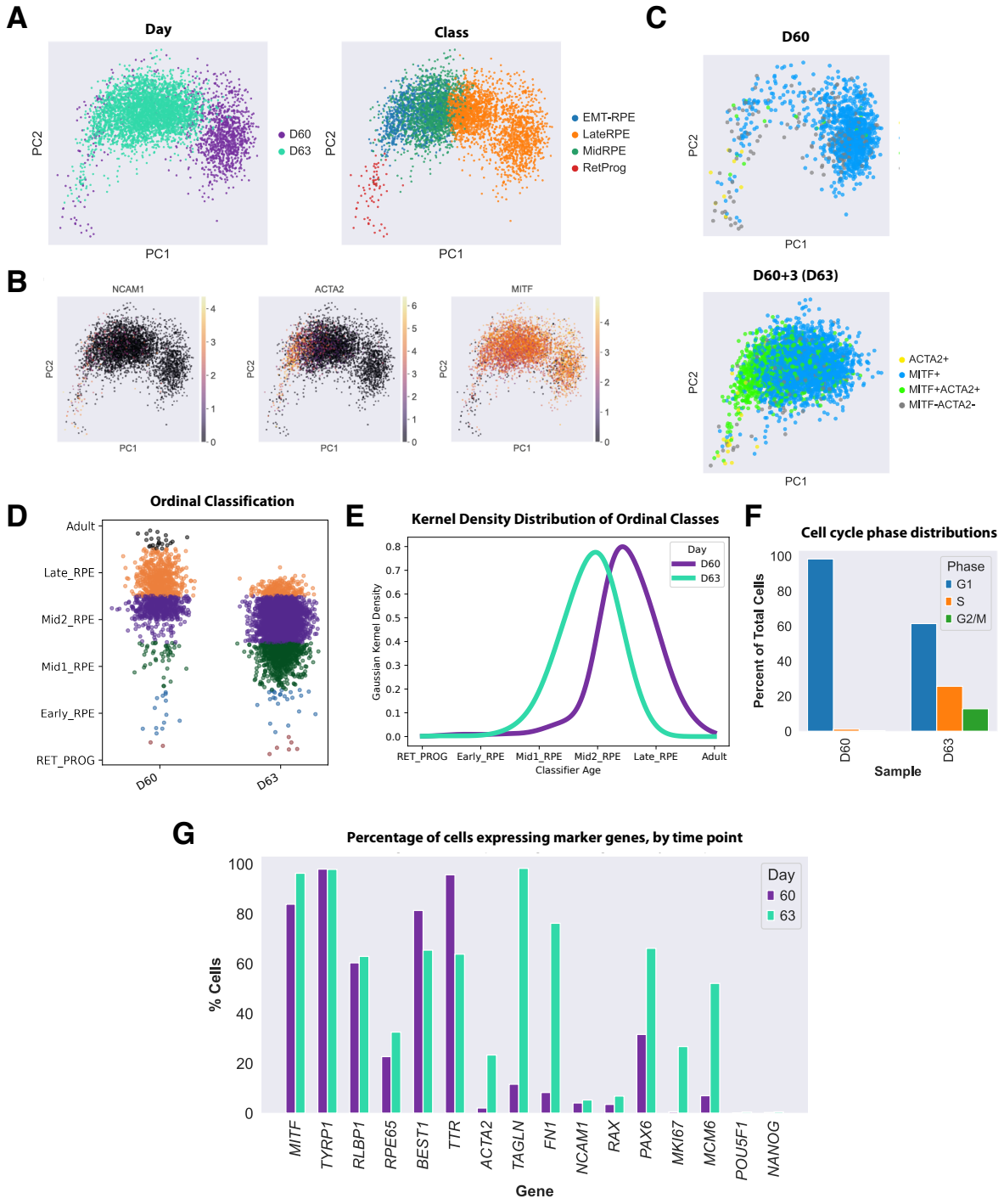


Figure A2.1. Single-cell transcriptomes of hESC-RPE after cryopreservation show a global de-differentiation and increased proliferative capacity. (A) Principal component (PC) representation of D60 (1,236 cells) and D63 (3,417 cells; D60+3 additional days after a second replating step) populations colored by time point (left) or cell type class (right). **(B)** Scatter plots of log2 normalized expression for neuroepithelial progenitor *(NCAM1)*, EMT-RPE *(ACTA2)*, and RPE *(MITF)* marker genes. **(C)** PC plots showing EMT status of D60 (top) and D63 (bottom) cells, according to the metrics previously defined in Fig. S2.5. **(D)** Dot plot showing ordinal classification of single hESC-derived KARO1 cells at D60 (1,236 cells) and D63 (3,417 cells) of differentiation at six different time points along embryonic stages (c.f. **Fig. 2.6F**). **(E)** Gaussian kernel density graph showing classification distribution for cells in (D) (c.f. **Fig. 2.6G**). Ordinal classification was performed as previously described in Petrus-Reurer, Lederer, et al (2022). **(F)** Bar plot showing the percentage of total cells at D60 and D63 in the proliferative (S, G2/M) and non-proliferative (G1) cell cycle phases. A larger fraction of cells are cycling at D63 than D60. **(G)**

Bar plot showing the percentage of D60 and D63 cells positively expressing marker genes of early RPE *(MITF, TYRP1, RLBP1)*, late RPE *(RPE65, BEST1, TTR)*, EMT-RPE *(ACTA2, TAGLN, FN1)*, retinal progenitor *(NCAM1, RAX, PAX6)*, proliferative *(MKI67, MCM6)*, and pluripotency *(POU5F1, NANOG)* cell types. A cell was defined as having positive expression if the normalized log2 gene expression was greater than 1. A larger fraction of D60 cells express late RPE markers, whereas more D63 cells express EMT-RPE and proliferative genes. There are no expressed pluripotency markers at both time points.

3. Statistical inference with a manifold-constrained RNA velocity model uncovers cell cycle speed modulations

Alex R. Lederer[a], Maxine Leonardi[b,*], Lorenzo Talamanca[b,*], Antonio Herrera[a], Colas Droin[b], Irina Khven[a], Hugo J.F. Carvalho[b], Alessandro Valente[a], Albert Dominguez Mantes[a,c], Pau Mulet Arabí[b], Luca Pinello[d,e,f], Felix Naef[b,†], Gioele La Manno[a,†]

(a) Laboratory of Brain Development and Biological Data Science, Brain Mind Institute, Faculty of Life Sciences, École Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland
(b) Laboratory of Computational and Systems Biology, Institute of Bioengineering, Faculty of Life Sciences, École Polytechnique Fédérale de Lausanne (EPFL), Switzerland
(c) Laboratory of Bioimage Analysis and Computational Microscopy, Institute of Bioengineering, Faculty of Life Sciences, École Polytechnique Fédérale de Lausanne (EPFL), Switzerland
(d) Molecular Pathology Unit, Massachusetts General Research Institute, Charlestown MA 02129, United States
(e) Massachusetts General Hospital Cancer Center, Harvard Medical School, Charlestown MA 02129, United States
(f) Broad Institute of MIT and Harvard, Cambridge MA 02139, United States

* These authors contributed equally

† Corresponding authors: gioele.lamanno@epfl.ch, felix.naef@epfl.ch

Currently under peer-review and available as a *bioRxiv* preprint (Lederer et al., 2024) at: https://doi.org/10.1101/2024.01.18.576093

I am first author of this research work.

3.0. Preface

In this chapter, I describe research carried out as a collaboration between the groups of Dr. Felix Naef and Dr. Gioele La Manno at the EPFL. The findings described here are adapted from the preprint version of a research article entitled “Statistical inference with a manifold-constrained RNA velocity model uncovers cell cycle speed modulations.” I am the first author of this publication. I developed the *VeloCycle* model, performed all computational analyses, designed the figures, and wrote the manuscript. Experimental data collection and

model conceptualization was performed in part with the help of other author contributions, as indicated in Chapter 3.6.4.

3.1. Synopsis

Across a range of biological processes, cells undergo coordinated changes in gene expression, resulting in transcriptome dynamics that unfold within a low-dimensional manifold. Single-cell RNA-sequencing (scRNA-seq) only measures temporal snapshots of gene expression. However, information on the underlying low-dimensional dynamics can be extracted using RNA velocity, which models unspliced and spliced RNA abundances to estimate the rate of change of gene expression. Available RNA velocity algorithms can be fragile and rely on heuristics that lack statistical control. Moreover, the estimated vector field is not dynamically consistent with the traversed gene expression manifold. Here, we develop a generative model of RNA velocity and a Bayesian inference approach that solves these problems. Our model couples velocity field and manifold estimation in a reformulated, unified framework, so as to coherently identify the parameters of an autonomous dynamical system. Focusing on the cell cycle, we implemented *VeloCycle* to study gene regulation dynamics on one-dimensional periodic manifolds and validated using live-imaging its ability to infer actual cell cycle periods. We benchmarked RNA velocity inference with sensitivity analyses and demonstrated one- and multiple-sample testing. We also conducted Markov chain Monte Carlo inference on the model, uncovering key relationships between gene-specific kinetics and our gene-independent velocity estimate. Finally, we applied *VeloCycle* to *in vivo* samples and *in vitro* genome-wide Perturb-seq, revealing regionally-defined proliferation modes in neural progenitors and the effect of gene knockdowns on cell cycle speed. Ultimately, *VeloCycle* expands the scRNA-seq analysis toolkit with a modular and statistically rigorous RNA velocity inference framework.

3.2. Introduction

Single-cell RNA-sequencing (scRNA-seq) captures a static snapshot of gene expression in a destructive manner, making it difficult to interpret dynamical aspects of biological processes. To address this issue, computational approaches have emerged that reconstruct temporal information among cellular states from scRNA-seq data (Lederer & La Manno, 2020). For example, RNA velocity exploits the ratio between unspliced and spliced transcripts to estimate a vector that describes the rate of change of gene expression (La Manno et al., 2018). The model considers a system of first-order ordinary differential equations describing the mRNA life cycle and whose key parameters are splicing and degradation rates.

Under simplified assumptions, it is possible to estimate these parameters from data (Svensson & Pachter, 2018).

The original RNA velocity framework, implemented in *velocyto*, fixes a common splicing rate across genes to infer a relative gene-dependent degradation rate from spliced-unspliced phase portraits (La Manno et al., 2018). This parameter is then plugged into the differential equations to obtain a gene-specific velocity. An extended model for the estimation of RNA velocity is the "dynamical model," implemented for the first time in the tool *scvelo*, which introduced for each gene a cell-wise latent time to support the estimation of kinetic parameters varying across a pseudotemporal axis, making them directly identifiable (Bergen et al., 2020). By exploiting expectation-maximization, *scvelo* estimates latent time and kinetic parameters. Other methods have harnessed these modeling ideas or worked towards extending them (Burdziak et al., 2023; Z. Chen et al., 2022; M. Gao et al., 2022; Gorin et al., 2020; Lange et al., 2022; C. Li et al., 2022; Qiao & Huang, 2021; X. Qiu et al., 2022; Tedesco et al., 2022; Weng et al., 2021). However, RNA velocity analysis remains highly sensitive to pre-processing choices and requires various heuristics to obtain the final estimates.

A pervasive yet potentially dangerous heuristic is the nearest-neighbor smoothing used to approximate expectations on the RNA counts; this procedure can let information bleed from some genes to others and cause distortions (Bergen et al., 2021). Additionally, the use of general non-linear dimensionality reduction techniques to bring the high dimensional velocity vector onto a two-dimensional embedding (e.g., UMAP, tSNE) risks introducing artifacts (Chari & Pachter, 2023). For instance, velocities associated with orthogonal processes, such as proliferation and differentiation, may be blended together, and adjacent yet unrelated cell populations might affect the resulting vector. Other algorithmic steps and corner cases that typically require attention have already been noted (Gorin et al., 2022; La Manno et al., 2018). However, a seldom discussed, yet central, limitation of most RNA velocity models is that velocity estimation is not performed jointly on all genes. This strategy is problematic, even when some form of global reconciliation is sought; for example, when aggregating individual latent times into a global one, the obtained kinetic parameter estimates remain independent. This leads to a physically and geometrically inconsistent velocity vector, whose gene-specific components are on different timescales and whose resulting direction is not necessarily tangent to the low dimensional manifold cells traverse. This is inappropriate for unbiased forecasting, as future states predicted by integration are bound to rapidly escape the gene manifold and inhabit unlikely regions of the expression space.

Finally, the lack of established ground truths for RNA velocity limits the rigorousness of sensitivity analyses that can be performed on newly developed methods, creating a

challenging environment to benchmark advanced extensions (Aivazidis et al., 2023; Gayoso et al., 2023; Gu et al., 2022; Qin et al., 2022). In particular, overparameterization becomes a concern, especially for models with less stringent assumptions, several non-linearities, or many degrees of freedom. Furthermore, proposed Bayesian formulations of the "dynamical model" return a high-dimensional mean-field posterior, which is not consistent with the assumption of low rank dynamics and is poorly suited to inference on the velocity and statistical comparisons of cell population dynamics.

We addressed these challenges by reformulating RNA velocity analysis as an inferential framework rooted in a manifold-constrained probabilistic model. Adopting this approach, we propose an explicit parametrization of RNA velocity as a field defined on the manifold coordinates. We focus on one-dimensional periodic manifolds in a framework called *VeloCycle*, enabling model validation and application to cell cycle dynamics. The cell cycle is the most ubiquitous periodic process in biology and plays a fundamental role in embryonic development, tissue regeneration, and disease (Tyson & Novák, 2022; Wiman & Zhivotovsky, 2017). Despite being pervasive in scRNA-seq datasets, default cell cycle analysis pipelines (Satija et al., 2015; Wolf et al., 2018) are still restricted to categorical phase assignment based on a small selection of marker genes (Eastman & Guo, 2020; Schwabe et al., 2020; Tirosh et al., 2016). In this work, we not only tackle the broader issue of maintaining geometrical constraints during velocity estimation, but also make strides in improving cell cycle analysis in scRNA-seq data, highlighting its continuous nature and providing control over the actual biological time scales. We apply *VeloCycle* across different biological contexts, experimentally benchmark against time-lapse microscopy measurements, and illustrate the ability to perform statistical tests.

3.3. Results

3.3.1. Manifold-constrained RNA velocity addresses shortcomings of other approaches

We first sought to redesign RNA velocity estimation by unifying manifold and velocity inference into a single probabilistic framework (**Fig. 3.1A, left**). This framework is articulated around a generative model with explicit low-dimensional dynamics at its core. In our model, cells move in time as points on a low-dimensional manifold x embedded within the space of all measured genes. Spliced and unspliced molecules are formulated as a function of x only (i.e., $s(x)$, $u(x)$). Then, by parameterizing velocity as a function of the manifold coordinates $V(x)$, we constrain RNA velocity vectors to lie tangent to the manifold (**Fig. 3.5.1A, right**). This is contrary to previous approaches where velocity direction is unconstrained, as it is the result

of gene-wise estimates (Bergen et al., 2021; Gorin et al., 2022) (**Fig. 3.1B**). We take the derivative of the expected spliced counts, apply the chain rule, and plug in the kinetic equations to obtain a velocity vector field interlocking the kinetic parameters of all genes and the dynamics of the latent coordinates (**Methods 3.5.1-3.5.3**). Noise in the measured raw read counts is modeled as a negative binomial, also as a function of the manifold *x*, and biochemically informed priors are chosen for all other parameters, including splicing (β) and degradation (γ) rates for each gene (**Fig. 3.1C; Methods 3.5.4**).

This formulation constitutes a latent variable framework for estimation of the gene expression manifold and RNA velocity. The choice of a specific dimensionality, topology, and associated functional parametrization constraining its geometry can be tailored in an application-specific manner (**Fig. 3.1D**). We propose inference in two statistical learning procedures: (1) *manifold-learning* to jointly learn the parameters defining the geometry of the gene expression space and assign each cell a manifold (latent) coordinate, and (2) *velocity-learning* to find a velocity field and kinetic parameters, conditioned on the manifold geometry and cell coordinates (**Fig. 1D-E**).

We implemented this scheme considering a scenario where the prior information on manifold topology is strong: the cell cycle, a one-dimensional periodic space on which gene expression varies smoothly and can be parametrized using a Fourier series. Our framework, *VeloCycle*, constitutes a generative probabilistic model with two groups of latent variables and is solved in Pyro (Bingham et al., 2018.) (**Methods 3.5.4, Table 3.1**). The first group relates to manifold-learning and defines the low-dimensional manifold *x* parameterized as cell cycle phase (ϕ) and gene-specific Fourier coefficients (v_0, v_{1sin}, v_{1cos}) using the expected spliced counts as a function of the phase (**Fig. 3.1E, S3.1A-B**). The second group relates to velocity-learning from the expected unspliced counts and includes the gene-specific degradation rates (γ_g), effective splicing rates (β_g) and velocity harmonic coefficients ($v\omega$), which parameterize an angular speed function ($\omega(\phi)$) describing how cell cycle velocity changes along the manifold (ϕ) (**Fig. 1E, S1C-D; Methods 3.5.4**). Using stochastic variational inference (SVI), *VeloCycle* returns the joint posterior probability of the latent variables, which can be used to (i) perform statistical velocity significance testing, (ii) characterize underlying correlations between the uncertainty of latent variables, (iii) estimate cell cycle velocities on a biologically-relevant time scale, and (iv) facilitate the application of velocity to small datasets by transfer learning (**Fig. 3.1F**).

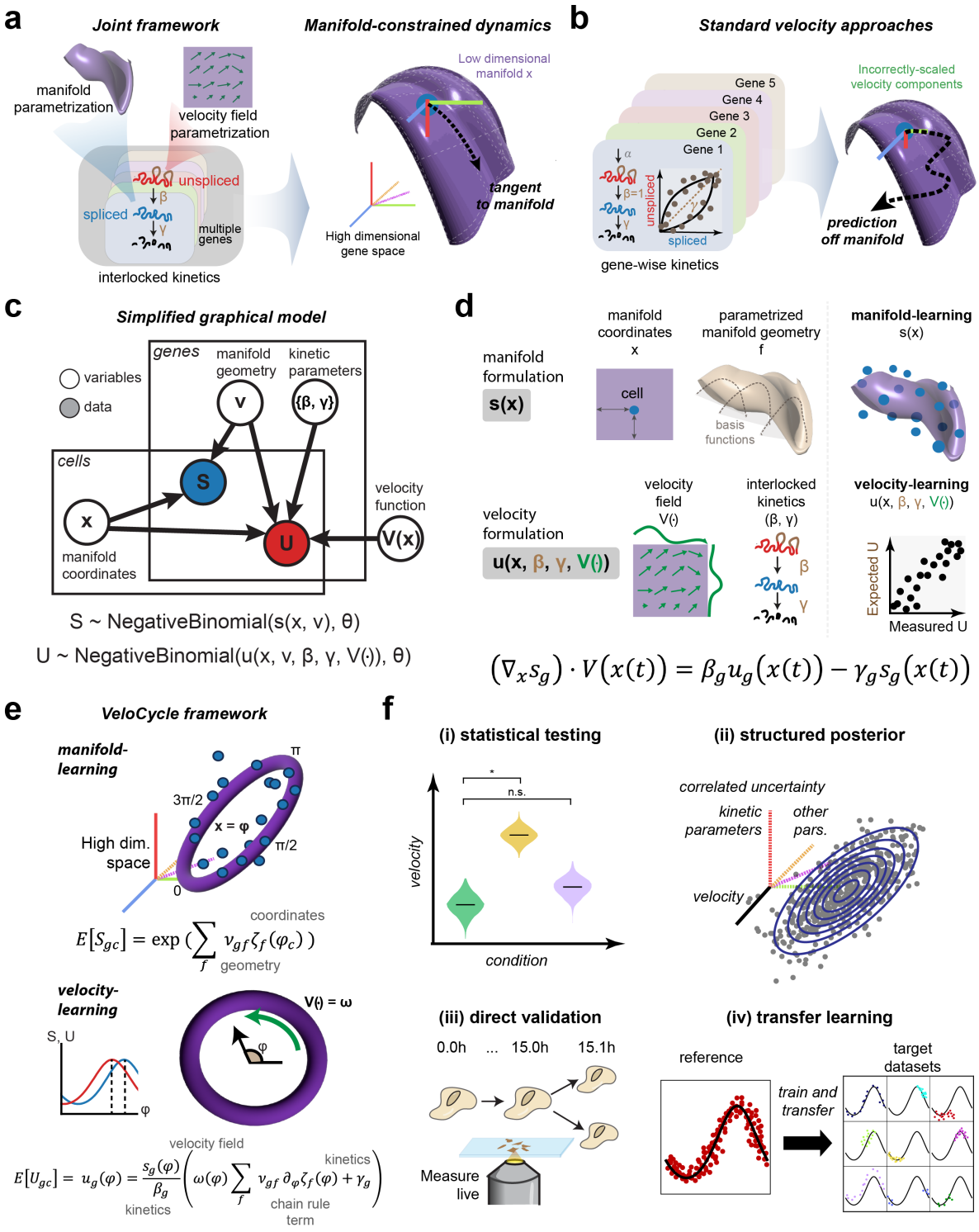


Figure 3.1. Statistical inference of RNA velocity with a manifold-constrained framework for the cell cycle. (A) Schematic of a joint framework for parameterization of the gene expression manifold and RNA velocity field. By defining velocity as a function of the manifold coordinates, the velocity vector field is constrained to be tangent to the manifold. This is achieved by interlocking the kinetic parameters of all genes with latent coordinate dynamics. **(B)** Schematic of unconstrained velocity estimation described by standard velocity approaches. By estimating the vector field as a combination of incorrectly-scaled, gene-dependent components, velocity is no longer tangent to the manifold. **(C)** Plate diagram of the probabilistic relationship among latent variables and observable data (S, U), modeled using a negative binomial distribution. S is sampled from the expectation, manifold coordinates, and

manifold geometry. U is sampled from the manifold information, kinetic parameters, and velocity function. **(D)** Top: manifold formulation is defined for the spliced counts (s) using the cell-specific manifold coordinates (x) and a gene-specific geometric family (f), with which observed data can be directly mapped to the high-dimensional manifold space. Bottom: velocity formulation is defined for unspliced counts (u) as a velocity field function (V) and interlocked kinetic parameters (β, γ). We can obtain a velocity estimate by taking the chain rule over these entities. **(E)** Schematic of the two procedure steps used by *VeloCycle* to solve manifold-constrained velocity estimation for periodic processes such as the cell cycle. First, *manifold-learning* estimates the manifold coordinates and geometry; second, *velocity-learning* estimates the kinetic parameters and manifold-dependent velocity function. **(F)** Schematic of some types of velocity analyses that are possible for the first time with *VeloCycle*, including: (i) statistical credibility testing between multiple samples and against a zero-velocity null hypothesis; (ii) posterior marginal distribution analysis of velocity and kinetic parameters by Monte-Carlo Markov Chain (MCMC) sampling; (iii) extrapolation of velocity to real biological time of cell cycle speed with live microscopy; and (iv) transfer learning of gene manifold from high-content, quality references to low-content, noisy target datasets.

3.3.2. Sensitivity analysis on simulated data validates *VeloCycle*

After designing our model, we sought to evaluate its performance on simulated data, as no real dataset is endowed with ground truth information for phases, speed, and RNA kinetic parameters. We employed a simulation intended to preserve important relations expected in real data (La Manno et al., 2018) and avoid biologically improbable scenarios (**Methods 3.5.5; S3.2A-C**). Specifically, we incorporated positive correlations among the splicing and degradation rates (r=0.30) and baseline expression levels (r=0.30) (**Fig. S3.2A**). This structure naturally imposed a positive correlation between the splicing rate and total spliced counts as well as a negative correlation between the splicing rate and total unspliced counts (**Fig. S3.2B-C**).

First, we evaluated *manifold-learning* across 20 individually simulated datasets each containing 3,000 cells and 300 genes and found *VeloCycle* inferred phases that closely matched the ground truth, with a circular correlation of $r_\phi = 0.95$ (**Fig. 3.2A-B**). The estimation error was consistently smaller than the uncertainty defined by the posterior, with true values falling within the 90% credible interval for 99.2% of cells (**Fig. S3.2D**). We also verified that the gene-specific Fourier series coefficients closely tracked the original ground truths (r_{v0} = 0.95, r_{v1sin} = 0.98, and r_{v1cos} = 0.98) (**Fig. 3.2C, S3.2E**). For these parameters, wider credible intervals corresponded to more noisy genes with a larger coefficient of variation (**Fig. S3.2F**). Overall, these results confirmed that *VeloCycle* correctly identified the manifold geometry and cell coordinates. To assess robustness of the model on different dataset sizes, we performed sensitivity analysis, varying the number of cells and genes (**see Methods 3.5.5**). We found that estimates were broadly accurate, with a circular correlation coefficient greater than 0.70 obtained using as few as 100 cells or 100 genes (**Fig. 3.2D**). We further benchmarked our

inference against DeepCycle, a recent autoencoder-based method (Riba et al., 2022). This comparison showed that *VeloCycle* was typically more accurate (60% lower MSE on average, $r_\phi = 0.95$) than DeepCycle ($r_\phi = 0.73$), despite the latter using velocity moments to achieve its estimations (**Fig. 3.2E-H**).

Next, we conditioned *VeloCycle* on the simulated phase and gene harmonics to assess *velocity-learning*. We observed accurate estimation of gene-wise kinetic parameters across 20 individually simulated datasets, with a particularly close match of degradation-splicing rate ratios to the ground truth ($r_{\gamma/\beta} = 0.997$, $r_\beta = 0.918$, $r_\gamma = 0.617$; **Fig. 3.2I, S3.2G-H**). Importantly, *VeloCycle* was capable of returning an accurate estimate of the mean angular velocity (percent error running 5.4-22.6%; **Fig. 3.2J**). *VeloCycle* recovered the biological correlation structure among estimated kinetic parameters and total counts, without imposing them in the model formulation (**Fig. 3.2K, cf. Fig. S3.2A-B**).

We performed sensitivity analysis to understand how the estimations behaved at different ground truth velocities. We considered a large span of cell cycle velocities fully encompassing the range of biologically plausible ones (16 values from 0 to 1.5 radians per mean half-life, or rpmh, four simulations each). The results highlighted a stable performance of the method, with estimates 0.2-35.8% away from the ground truth (**Fig. 3.2L-M**). Error increased at slower velocities, with a lower Pearson's correlation between kinetic parameters and ground truths (**Fig. S3.2I, left**). Indeed, slower velocities corresponded to shorter delays between unspliced and spliced RNAs (**Fig. 3.2N; Methods 3.5.4.5**), which are more difficult to characterize accurately. In all simulations, the degradation-splicing rate ratios almost perfectly matched the ground truth (mean $r_{\gamma/\beta}$=0.99) (**Fig. S3.2I, right**). Finally, we investigated whether *velocity-learning* performance was affected by dataset size. We detected a dependence on the number of cells and genes, with the highest accuracy and tightest posterior ranges obtained on larger datasets; however, using more cells could compensate for fewer genes, and vice versa (**Fig. 3.2O and S3.2J**). We established 500 cells (and a minimum of 50 genes) or 350 genes (and a minimum of 50 cells) as the lower limits at which accurate velocity estimation can be performed.

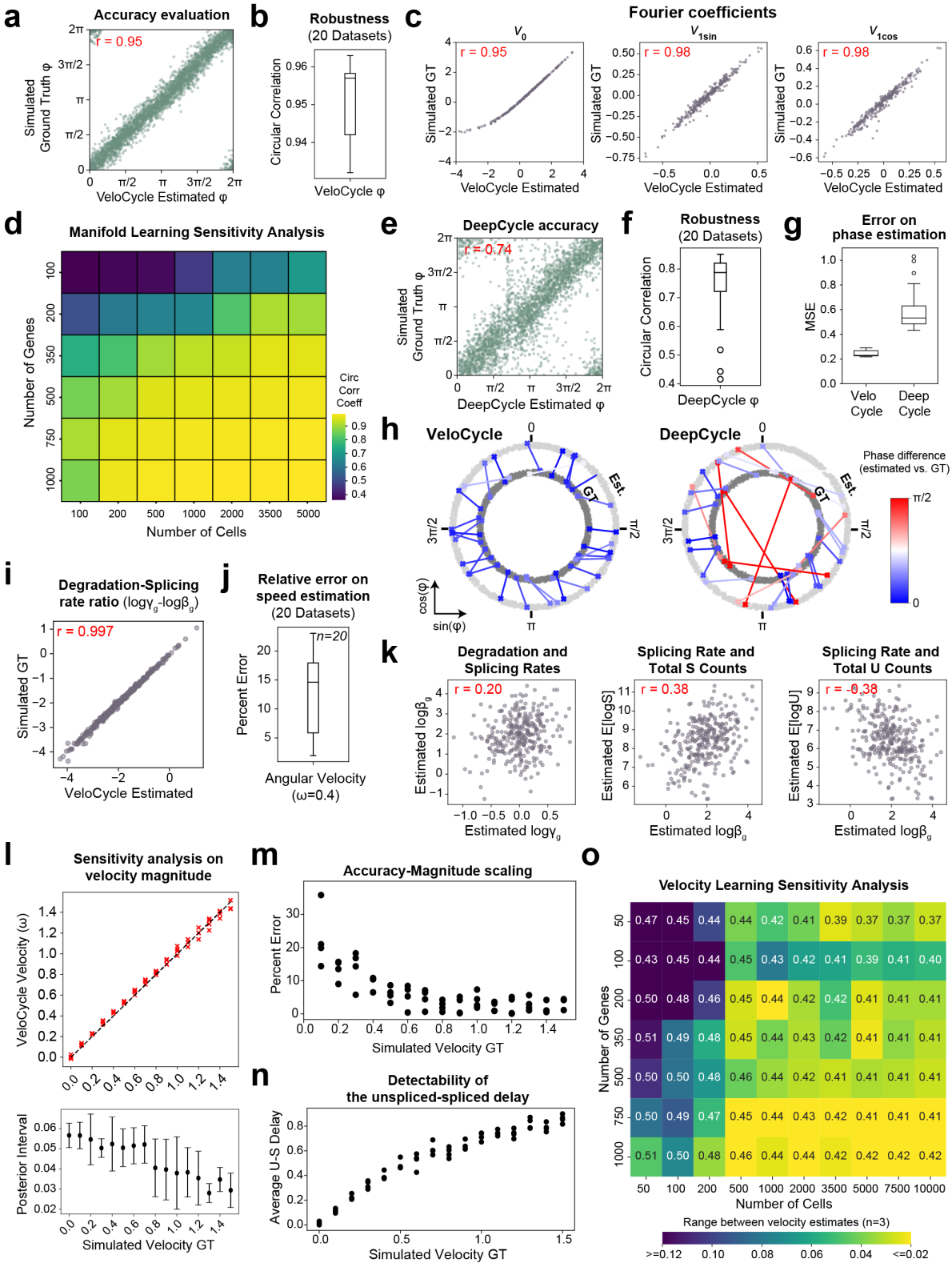


Figure 3.2. Sensitivity analysis of *VeloCycle* on simulated data. (a) Scatter plot of cell cycle phase assignment (VeloCycle Estimated) compared to the simulated ground truth (GT). **(b)** Box plot of circular correlation coefficients between estimated and GT phases across 20 independently simulated datasets, each containing 3,000 cells and 300 genes. **(c)** Scatter plots of estimated and GT values for the gene harmonic coefficients (v_0, v_{1sin}, v_{1cos}) using the dataset in (a). **(d)** Heatmap of the mean circular correlation coefficient between estimated and GT phases computed with varying numbers of cells and

genes. Each value is the average of three independent simulations. **(e)** Scatter plot of cell cycle phase estimation obtained by DeepCycle (Riba et al., 2022) compared to simulated ground truth. The same dataset shown in (a) was used. **(f)** Box plot of circular correlation coefficients between DeepCycle-estimated and GT phases across the datasets shown in (b). **(g)** Box plots of per-cell mean squared error (MSE) for phase estimation with *VeloCycle* and DeepCycle. **(h)** Polar plots representing the phase difference between estimated and simulated GT for 30 randomly chosen cells from a single simulated dataset using *VeloCycle* (left) and DeepCycle (right). Each dot represents a cell, and lines connect the estimated phase assignment (Est; light gray) to simulated ground truth (GT; dark gray). **(i)** Scatter plot of estimated ratio between γ_g and β_g compared to simulated GT for 300 genes. **(j)** Box plot of percent error between estimated and GT velocity (ω) across 20 simulated datasets with a GT of 0.4. **(k)** Scatter plots illustrating the recovered relationships among splicing rate ($\log\beta_g$), degradation rate ($\log\gamma_g$), spliced counts, and unspliced counts for 300 simulated genes. **(l)** Top: scatter plot of estimated and GT estimates for 16 different simulated velocities between 0.0 to 1.5 radians per mean half-life (rpmh) for 4 independently-simulated datasets. Bottom: box plots of posterior uncertainty intervals corresponding to the above simulations. **(m)** Scatter plot of percent error between estimated and GT velocity across conditions in (l). **(n)** Scatter plot of mean unspliced-spliced expression delay across conditions in (l). **(o)** Sensitivity analysis heatmap of the range among velocity estimates for 3 independently-simulated datasets, using varying numbers of cells and genes. The text value in each box represents the mean velocity over the 3 datasets, and intensity of the heatmap represents absolute range. The Pearson's correlation coefficient (r) over 20 individual simulated datasets is indicated in red in (a), (c), (e), (i), and (k). Each green dot represents a single gene in (a) and (e). Each purple dot represents a single gene in (c), (i), and (k).

3.3.3. *VeloCycle* manifold-learning estimates accurate and robust phases

After validating on simulated data, we deployed *VeloCycle* on real datasets produced with different scRNA-seq chemistries. We reasoned that access to a cell cycle phase ground truth, even if categorical (e.g., G1, S, G2/M), would facilitate the evaluation of our phase assignments. Thus, we performed *manifold-learning* on a Smart-seq2 dataset of fluorescence ubiquitination-based cell-cycle indication (FUCCI) system-transduced mouse embryonic stem cells (mESC) that were index-sorted using fluorescence-activated cell sorting (FACS) (Buettner et al., 2015). We fit the cell cycle phase on spliced counts using a gene set representing a broad gene ontology (GO) query (Ontology Consortium et al., 2023) (**Methods 3.5.4.3, 3.5.1**) and evaluated the results against FUCCI-FACS categories. Cells belonging to the same category were assigned to similar phases (**Fig. 3.3A-B**); a classifier based on two thresholds and trained on *VeloCycle* phases achieved 82.7% accuracy in predicting the annotations, almost matching the 87.8% accuracy obtained when training a logistic classifier on all genes (**Fig. 3.3C**). Furthermore, gene fits underlying *manifold-learning* closely replicated the expected sequential patterns of cell cycle genes. Among fits of high confidence were early-peaking histone acetylase *Hat1*, followed by transcription factor *Trp53*, and the later anaphase-promoting complex member *Ube2c* (**Fig. 3.3D**). The gene succession and oscillation amplitude were recapitulated when performing *manifold-learning* on a smaller set

of 209 genes, anticipating that the method is effective on chemistries with lower sensitivity (**Fig. 3.3E-F**).

Given our Fourier parametrization, we could classify genes by the phase of peak expression, oscillation amplitude, and estimation uncertainty (**Table 3.2**). Inspection of phase-amplitude relations revealed that marker genes typically used for scoring in packages such as Seurat and scanpy (Satija et al., 2015; Wolf et al., 2018) (henceforth "standard markers") clustered by phase, consistent with the FACS-based ground truth (**Fig. 3.3G-H**). Compared to non-markers, standard markers on average had a higher amplitude (mean 0.14 versus -0.15) and lower posterior uncertainty (standard deviation 0.26 versus 0.43) (**Fig. 3.3G**). However, of the top 200 periodic genes based on amplitude, the vast majority (74.5%) were not standard markers (**Fig. 3.3H**), and many (n=78) could be equally or more confidently trusted (i.e., tighter posterior probability) as cell phase predictors (**Fig. 3.3I**). Among those were calcium-binding protein *Calm2*, splicing co-factor *Son*, and cyclin *Ccnb1*, which all play roles in cell proliferation (Berchtold & Villalobo, 2014; Gruber et al., 2019; Jeon, 2013; A. Sharma et al., 2010).

We continued our scrutiny of *manifold-learning* using 10X Chromium data of human fibroblasts (**Fig. 3.3J-K; Table 3.3**). To put *VeloCycle* in relation to other approaches, we compared its estimated phases to those obtained by DeepCycle (Riba et al., 2022), finding a strong correspondence (human fibroblasts: r=0.882; **Fig. 3.3L**). Therefore, *VeloCycle* accomplishes similar phase estimation to DeepCycle but without using velocity and in tandem with fitting individual gene harmonics. As further validation that the correct cell cycle dynamics were captured, we observed a gradual increase in total UMIs along the phase, followed by a sharp drop corresponding to cytoplasm partitioning during cytokinesis (**Fig. 3.3M)**. These results highlight that *manifold-learning* estimates a biologically-meaningful one-dimensional geometric space that tracks with the cell cycle across scRNA-seq chemistries.

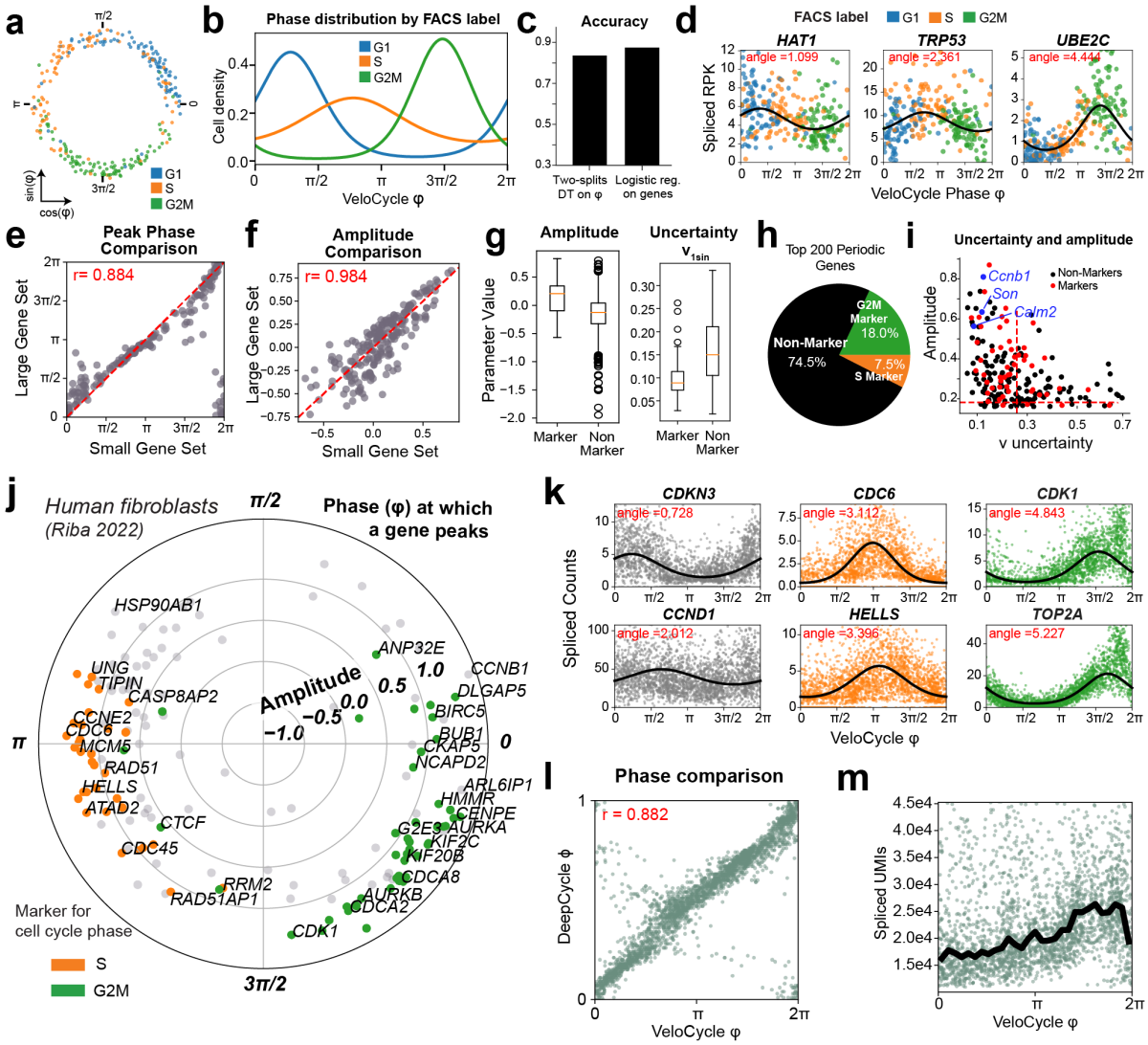


Figure 3.3. ***Manifold-learning*** **and gene periodicity on different datasets and technologies.** **(a)** Scatter plot representing the phase assignment of 279 mouse embryonic stem cells, colored by their FACS-sorted categorical phase (G1, S, G2/M) (Buettner et al., 2015). **(b)** Density plot for FACS-sorted labels (G2, S, G2M) across the phase assigned by *VeloCycle*. **(c)** Bar plot reporting categorical phase predictor obtained using a two-thresholds decision tree trained on the *VeloCycle* phase estimates only versus a logistic regression classifier trained on the entire gene expression matrix. **(d)** Representative scatter plots of genes fits. Curved black lines indicate a gene-specific Fourier series obtained with *manifold-learning*. The "peak" indicates the position of maximum expression along the cell cycle manifold (VeloCycle ϕ). **(e)** Scatter plot of gene-wise peak position of maximum expression using a small (x-axis) or large (y-axis) gene set during *manifold-learning* for FACS-sorted mESC data. **(f)** Scatter plot of gene-wise peak position of gene-wise amplitude using a small or large gene set. **(g)** Box plots of gene-wise amplitude and harmonic coefficient uncertainties for marker and non-marker genes for FACS-sorted mESC. **(h)** Pie chart of categorical composition for the top 200 periodic genes, as determined by amplitude. **(i)** Scatter plot of gene-wise total harmonic coefficient (v) uncertainty and amplitude. Gene dots are colored as standard "markers" or "non-markers". Red dashed lines represent the mean values for "markers." **(j)** Polar plot of estimated gene harmonics for human fibroblasts data (Riba et al., 2022). Each dot represents a gene (n=160). The position along the circle represents the phase of maximum expression, and the distance from the center represents total amplitude. Colored genes (orange/green) are those used to compute a standard cell cycle score with *scanpy* or Seurat. **(k)** Selected scatter plots of genes fits for markers of early (*CDKN3, CCND1*), mid (*CDC6, HELLS*), and

late (*CDK1, TOP2A*) cell cycle progression obtained for the human fibroblasts data. **(l)** Scatter plot of phases estimated with *VeloCycle* compared to DeepCycle for 2,557 human fibroblasts. The circular correlation is indicated in red. **(m)** Scatter plot of total raw spliced UMI counts by *VeloCycle* phase. Black lines indicate the binned mean UMI level.

3.3.4. Unspliced-spliced delays along *VeloCycle* phase identify realistic cell cycle velocities

We next investigated whether the unspliced molecule counts together with the *VeloCycle* phase are sufficiently informative to estimate cell-cycle velocity. To explore this intuitively before performing the full inference, one can extract phases and gene harmonics with *manifold-learning* for unspliced and spliced UMIs independently and use an approximate formula for the velocity that we derived (**Methods 3.5.4.2, 3.5.7**). We applied this approach on two cultures of human RPE1 cells that were grown in parallel and under identical conditions so that we could also assess robustness by replicate comparison. First, we extracted the phases on each of the datasets by *manifold-learning*, then we measured the delays (i.e., the phase difference) between peak unspliced and spliced expression for each gene (**Fig. 3.4A**). We observed consistent and positive delays for the genes (**Fig. 3.4B**) that correlated well between replicates (r=0.90; **Fig. 3.4C**). We interpreted this correlation as the first evidence that the data contains velocity information on the cell cycle, so we proceeded to estimate a cell cycle period with the aforementioned approximate formula. The calculation returned a period 18.5 times the average half-life, which corresponds to 18.5h assuming a realistic average half-life of 1h (**Fig. 3.4D**). In addition to being an approximation, another limitation of the point estimate is that it is not based on a proper noise model and is not associated with an uncertainty measure. To obtain a more accurate estimate and statistical measures of confidence, we learned the complete Bayesian model (*velocity-learning*) on both RPE1 replicates, conditioning on the random variables inferred by *manifold-learning*. Scaling the obtained velocity by the fitted average half-lives yielded average cell cycle periods of 20.1h ± 0.2h and 20.0h ± 0.2h (mean ± 95% credible intervals) for the two replicates (**Fig. 3.4E**). The posterior distributions broadly overlapped (71.2% overlap), indicating no credible velocity difference between the two replicates. To confirm on real data that *VeloCycle* can estimate cell cycle speed along a dynamic range relevant biologically, we performed *velocity-learning* on mESC, a rapidly-cycling cell type (Bertels et al., 2021; Eastman et al., 2020). For this dataset, *VeloCycle* returned an estimation of 10.5 ± 0.3 average half-life (**Fig. S3.3A**). As with RPE1 cells, the model recovered kinetic parameters with expected relationships among total UMI counts and gene-specific splicing and degradation rates, as previously observed in

simulated data (**Fig. S3.3B-E, cf. Fig. 3.2I**). Taken together, these findings confirm *VeloCycle* can estimate a cell cycle velocity and sample informative posterior distributions.

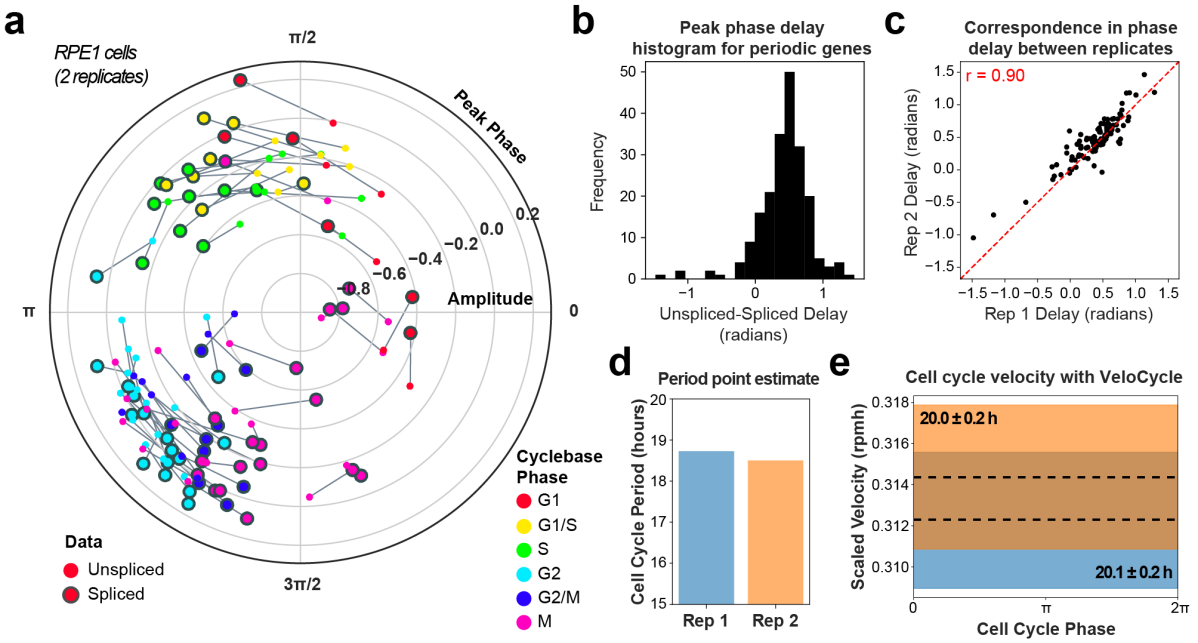


Figure 3.4. Analysis of delays and velocity scale in RPE1 cells. (**a**) Polar plot of the peak unspliced and spliced expression for 106 marker genes across two scRNA-seq replicates of RPE1 cells (4,265 and 9,994 cells) analyzed with *manifold-learning*. Genes are colored by their categorical annotation in Cyclebase 3.0 (Santos et al., 2015). Unspliced gene fits were inferred separately, conditioned on cell phases obtained when running *manifold-learning* on the spliced UMIs. (**b**) Histogram of unspliced-spliced delays (in radians) for 106 genes. Pearson's correlation is indicated in red. (**c**) Scatter plot of unspliced-spliced delays between two RPE1 cell line replicates. (**d**) Bar plot of the cell cycle periods obtained with a first-order-approximate point estimate (**see Methods**). (**e**) Posterior estimate plot of constant, scaled cell cycle speed (radians per mean half-life, rpmh) in two RPE1 cell line replicates. The black dashed lines indicate a mean of 500 posterior predictions, and the colored bar indicates the credibility interval (5th-95th percentile).

3.3.5. A structured variational distribution preserves uncertainty correlations and leads to better uncertainty estimates

Although we showed our variational formulation recovers accurate estimates of cell cycle phase and velocity in simulated and real data using stochastic variational inference (SVI), it is reasonable to question the limits of a simplified mean-field variational family in representing the structure of joint uncertainty among latent variables. We hypothesized that such a parametrization choice may lead to an overconfidence in the estimated velocity posterior because uncertainties on these latent variables may be inherently correlated (**Fig. 3.5A**). A piece of evidence in this direction was the observation that estimates on random gene subsets fell outside the posterior credible interval of the fit on all genes (**Fig. 3.5B**). To

eliminate this bias towards the underestimation of velocity uncertainty, we decided to characterize the model joint posterior by sampling it with Markov Chain Monte Carlo (MCMC; **Methods 3.5.4**). Using a No U-Turn Sampler, we studied the posterior for human fibroblasts (Riba et al., 2022), with MCMC revealing a five-times wider uncertainty compared to mean-field SVI (0.10 rpmh vs 0.02 rpmh; **Fig. 3.5C**).

Consistent with our hypothesis, this wider credible interval manifested along with a correlated joint posterior, capturing dependencies among the uncertainty of different latent variables. Specifically, examining the posterior, we found samples of the angular speed ($v\omega$) and degradation rate ($\log\gamma_g$) for certain genes that exposed a correlation structure (mean r = 0.26; **Fig. 3.5D**). Moreover, for each gene we noticed a strong correlation (mean r = 0.96) between posterior samples of splicing ($\log\beta_g$) and degradation ($\log\gamma_g$) rates (**Fig. 3.5E**). Both features cannot be captured by a mean-field variational distribution.

These findings advocated for a recrafting of our variational distribution to accommodate typical features of the posterior inferred by MCMC, in order to maintain inferential accuracy but avoid significantly time-consuming sampling procedures. We reformulated our variational distribution with $\log\gamma_g$ and $v\omega$ modeled as a low rank multivariate normal (LRMN) and with the $\log\beta$ for each gene modeled as a normal conditional on the corresponding $\log\gamma$ (**Methods 3.5.3.2**). Upon retraining this new SVI+LRMN model, we obtained a velocity estimate with a larger uncertainty range (0.08 rpmh) than with mean-field SVI (**Fig. 3.5C-E**). Additionally, we detected a correlation among the SVI+LRMN posterior samples between $\log\gamma_g$ and $v\omega$ for a subset of genes that overlapped with the results of MCMC; this resulted in a decreased Kullback-Leibler (KL) divergence between the SVI+LRMN and MCMC posteriors than between the SVI and MCMC posteriors (**Fig. 3.5F, S3.4A**).

Importantly, there was a correspondence between the specific genes with high $\log\gamma_g$ and $v\omega$ uncertainty correlation in both SVI+LRMN and MCMC (**Fig. S3.4B**). Genes with a greater correlation between $\log\gamma_g$ and $v\omega$ tended to be those with larger unspliced-spliced delay (**Fig. S3.4C**). We speculated the degree of dependence between a gene's $\log\gamma_g$ and $v\omega$ is related to the extent it contributes to the velocity estimate. This was supported by a leave-one-out experiment, where individual genes with smaller degradation rates were those most strongly affecting velocity estimates (**Fig. S3.4C-D**). The correlation between $\log\gamma_g$ and $v\omega$ posterior uncertainty was also reproducible when SVI+LRMN was applied to mouse ESCs (**Fig. S3.4E-F**). Overall, these implementation changes led to generation of a more robust model that can be confidently used for inference, while preserving the underlying correlation structure of the true posterior.

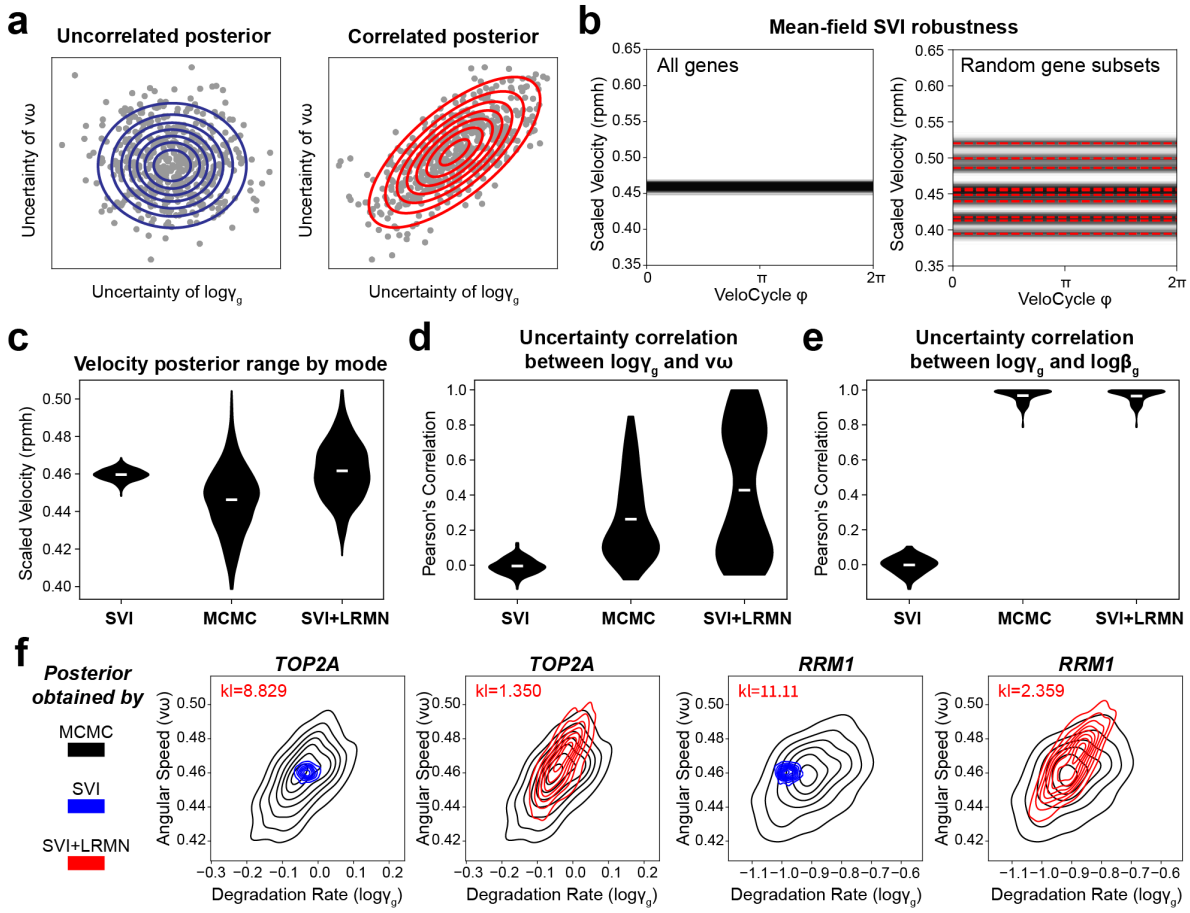


Figure 3.5. Relationships between parameter uncertainties and choice of the variational distribution. (**a**) Schematic of the hypothetical scenarios where a gene has uncorrelated (left) and correlated (right) posterior uncertainty between $\log\gamma_g$ and $v\omega$. Blue circles represent the Gaussian kernel density of the distribution, and red lines represent an uncertainty interval between two arbitrary fixed points. (**b**) Left: posterior estimated velocity plot inferred for 2,557 cultured human fibroblasts (Riba et al., 2022) using the original stochastic variational inference (SVI) mode of *VeloCycle*. Right: posterior estimated velocity plot on human fibroblasts, estimated with *VeloCycle* for ten random subsamples of the data, each using only 50% of the genes. (**c**) Violin plots of scaled velocity (in rpmh) for human fibroblasts after estimation using the stochastic variational inference (SVI), Monte Carlo Markov Chain (MCMC), and low rank multivariate normal (SVI+LRMN) *velocity-learning* modes. (**d**) Violin plots of Pearson's correlations between the degradation rate ($\log\gamma_g$) and angular speed ($v\omega$) posterior uncertainties across 160 genes using different *VeloCycle* modes. **(e)** Violin plots of Pearson's correlations between the degradation ($\log\gamma_g$) and splicing ($\log\beta_g$) posterior uncertainties across 160 genes using different *VeloCycle* models. (**f**) Density representation of overlapping $\log\gamma_g$-$v\omega$ posterior distributions between MCMC and either SVI (top) or SVI+LRMN (bottom) for *TOP2A* and *RRM2* (black: MCMC; blue: SVI; red: SVI+LRMN). Kullback-Leibler divergence scores are shown in red. All posterior means were taken over 500 predictive samples.

3.3.6. Cell tracking and labeling experiments validate computationally inferred velocities

Estimates of a manifold-constrained cell cycle speed with *VeloCycle* are most conveniently expressed in units of mean half-lives (i.e., gene degradation rates, **see Methods**

3.5.4.1, 3.5.4.6). Since the average values of half-lives are typically known in many cell types, real time estimates of RNA velocity can be obtained and validated along the cycle. In this respect, we reasoned that time-lapse microscopy offers a compelling means for comparing *VeloCycle* estimates to a ground truth.

To benchmark our velocity estimation framework against an experimentally-determined cell cycle period, we examined a dataset of dermal human fibroblasts (dHFs) monitored by time-lapse microscopy and for which scRNA-seq data was collected (Capolupo, Khven, et al., 2022) (**Methods 3.5.6.3**). Our SVI+LRMN model inferred a constant cell cycle period of 15.3 ± 1.2h, assuming an average half-life of the modeled transcripts of 1h (**Fig. 3.6A, S3.5A-D**). Next, we used *VeloCycle* to infer a non-constant (periodic) cell cycle velocity, and we obtained a similar estimated duration of 16.5 ± 2.1h, with maximal velocity near mitosis (approximately $3\pi/2 < \phi < 2\pi$) (**Fig. 3.6B**). We then reconstructed the cell cycle period using cellpose (Stringer et al., 2021) and TrackMate (Ershov et al., 2022) for 268 individual cells followed by time-lapse imaging (**Fig. 3.6C**). From these data, we recovered a median cell cycle of 15.8h (s.d. 3.1h), which overlapped with the posterior credibility interval of the *VeloCycle* estimate (**Fig. 3.6D, c.f. Fig. 3.6A-B**). Comparable results were obtained when using the smaller set of cycling genes (Riba et al., 2022) (**Fig. S3.5E**). Taken together, these results indicate an ability to obtain comparable cell cycle speed estimates from live-microscopy and *VeloCycle*.

We next stratified velocity by an independent categorical cell cycle phase to gain further granularity on these evaluations and model behavior. We observed a faster progression through the cell cycle during G2/M phase (mean scaled velocity of 0.47 rpmh) compared to a slower progression during G1 (0.37 rpmh) and S (0.36 rpmh) phases (**Fig. 3.6E**). Kinetic parameters and their posterior uncertainties were strongly correlated between the constant and periodic velocity models (**Fig. S3.5F-G**). Interestingly, when estimating the average unspliced-spliced delay for genes peaking at different cell cycle phases, we found that cell cycle phases with larger average delays corresponded to regions with faster velocity (**Fig. 3.6F**). Genes with larger delays were also those with smaller splicing and degradation rates, which is expected from the approximate model (**Fig. S3.5H; Methods 3.5.4.2**). After examining the unspliced-spliced delay and the low-rank gene-wise posterior correlation between the angular speed and degradation rate, we could identify specific genes that most strongly contributed to the underlying velocity estimates (**Fig. 3.6G**).

To further scrutinize the degree to which cell cycle durations inferred by *VeloCycle* match those obtained experimentally, we performed time-lapse microscopy and scRNA-seq on the same cultured RPE1 cells. The speed obtained with *VeloCycle* was approximately 17.7

± 2.1h (**Fig. 3.6H, Methods 3.5.7.3**); as in dHFs, this computational estimate overlapped with the mean cell cycle duration of 17.7h (standard deviation of 3.4h) obtained from tracking dividing cells by time-lapse imaging (338 cells) (**Fig. 3.6I-J**). We next sought to compare our cell cycle duration measurements from time-lapse microscopy and *VeloCycle* to those obtained using an orthogonal experimental technique. Therefore, we performed continuous EdU labeling to independently estimate cell cycle length (**Fig. 3.6K-L**). After monitoring EdU levels at 13 time points over 72 hours (**Fig. 3.6M**), we used p21 (CDKN1A) staining to account for cells in G0 and determined a mean cell cycle length of 16.8h (**Fig 3.6N-O; Methods 3.5.7.4**). Taken together, these findings validated the computational RNA velocity estimates in the context of the cell cycle. To our knowledge, this is the first example of a direct validation of RNA velocity estimation with experimental methodologies and justifies the use of *VeloCycle* output in units of real (i.e., no pseudo-) time.

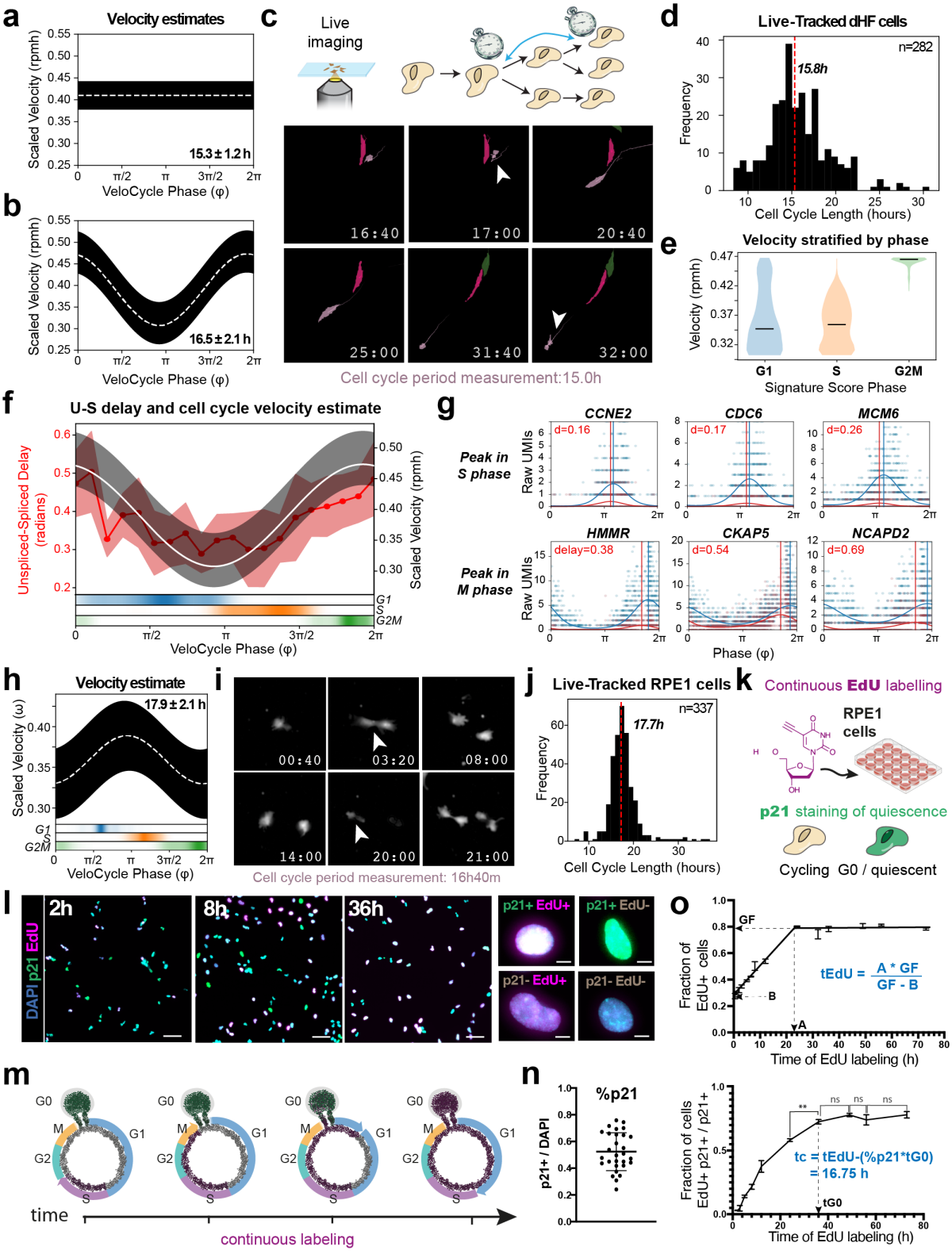


Figure 3.6. Validation of computationally inferred velocities by cell tracking and labeling experiments. (a) Posterior estimate plot of constant cell cycle speed in 1,222 dermal human fibroblasts (dHFs) (Capolupo, Khven, et al., 2022). **(b)** Posterior estimate plot of periodic (non-constant) cell cycle speed in cells from (a). **(c)** Top: schematic of time-lapse microscopy with live-imaging to track consecutive cell divisions. Bottom: example microscopy images at multiple time points to illustrate tracking a single segmented dHF cell (pink) through two divisions. Following division of the mother cell

(16:40h), one daughter cell is tracked for 15 hours until dividing itself (31:40h). **(d)** Histogram of cell cycle period for 268 dHFs tracked by live-microscopy. **(e)** Violin plot of cell cycle speed for dHFs, stratified by categorical phase assignment. **(f)** Dual-axis plot of the correspondence between the unspliced-spliced expression delay (left) and cell cycle velocity estimate (right). Left: genes were grouped in 20 equal bins by phase and the unspliced-spliced delay was calculated. Right: the scaled velocity estimate from (c). Bottom: cell cycle categorical (G1, S, G2/M) phase assignment probability. **(g)** Gene expression scatter plots for genes peaking in S (top) and M (bottom) phases. Vertical lines correspond to the peak phase of spliced (blue) and unspliced (red) counts, used to compute an unspliced-spliced delay in (f). **(h)** Posterior estimate plot of periodic (non-constant) cell cycle speed in 3,354 retinal pigmented epithelial cells (RPE1). **(i)** Representative microscopy images tracking a single RPE1 cell from birth (3:20h) to subsequent division (20:00h). **(j)** Histogram of cell cycle period for 337 RPE1 cells tracked by live-microscopy. **(k)** Diagram of the cumulative EdU/p21 experiment. Cells were continuously exposed to EdU, fixed at different timepoints, and subjected to EdU detection and p21 immunostaining. **(l)** Left: representative images of p21 (green), DAPI (cyan), and EdU (magenta) staining after cumulative EdU labeling for 2h, 8h, and 36h. Scale bar is 100um. Right: representative images of individual cells with different staining combinations. Scale bar is 10um. **(m)** Schematic of cumulative EdU labeling during cell cycle progression. Cycling cells incorporate EdU (magenta) when they undergo DNA replication (S phase). Thus, the duration of the EdU pulse is directly proportional to the fraction of EdU-positive cells. G0/quiescent cells are represented in green. The total number of cells consists of the number of cells in all cell cycle phases and G0, so the total duration of the cell cycle must be accounted for while excluding quiescent cells. **(n)** Dot plot representing the average percentage of p21+ cells along the different time points. Error bar indicates the standard deviation (SD), and each dot represents the percentage of p21+ cells for a single replicate (n=29). **(o)** Top: line plot of the fraction of EdU+ cells after at 13 time points (from 30 min to 73h). Data show the mean of three replicates (except for 2h, which is from two), and error bars indicate the SD. Accumulation of EdU can be divided into a linear growth phase and a plateau phase. Using quantities derived from a linear fit of the growth phase, we can derive a formula for calculating the time for total EdU (tEdU) labeling. Bottom: line plot of fraction of EdU-positive cells among quiescent cells (p21+) plotted as a function of time. The time when the fraction of EdU-positive cells among the quiescent population stops growing significantly is taken as an estimation of the tG0. The red dashed line indicates the median in (d) and (j). The white dashed line indicates the mean of 500 posterior predictions and the black bar indicates the credibility interval (5th-95th percentile) in (a), (b), and (h).

3.3.7. *VeloCycle* enables direct statistical velocity comparisons in response to drug treatment

Existing frameworks for RNA velocity do not propose an approach to test the statistical significance of obtained estimates, likely because it is challenging given a gene-wise velocity parametrization. For example, it is currently not possible to determine whether RNA velocity estimates close to zero should just be interpreted as noise. Furthermore, direct comparisons between velocity estimates of two samples cannot be supported by a measure of confidence. With *VeloCycle*, statistical inference testing on velocity is possible for the first time, both against a specific null-hypothesis and for differential velocity significance between cell populations.

To illustrate how our model can be used for statistical velocity tests in practice, we conducted RNA velocity analysis on a PC9 adenocarcinoma cancer cell line before (D0) and

after (D3) treatment with the drug erlotinib (Aissa et al., 2021) (**Fig. S3.6A-E**). Statistical testing in a Bayesian setting can be achieved by calculating credible intervals from the posterior. First, we considered the velocity posterior of the untreated (D0) cells to ask whether there is statistical support for a non-zero velocity. Given no overlap between the credible interval and zero, we could conclude the data contains statistically significant evidence for progression through the cell cycle (**Fig. 3.7A, left**). We then compared the treated sample (D3), with the control (D0). We found significant velocity differences between the D0 and D3 time points, where a slower mitotic cell cycle speed was detected at D3 (**Fig. 3.7B**). Such testing can be done globally and also locally. For example, we stratified by phase intervals and inspected the posterior samples, confirming a decreased speed during G2/M phase at D3 compared to D0, but not during G1 and S (**Fig. 3.7B, Fig. S3.6F**). The reduced presence of cells in M phase after erlotinib treatment was further suggested by the low density of D3 cells with assigned the phase coordinate (**Fig. S3.6A, bottom**).

Since the unspliced-spliced delay is linked with cell cycle velocity, we hypothesized there would be differential delays between the D0 and D3 time points, particularly for genes peaking during M phase. After calculating the gene-wise unspliced-spliced delay before and after erlotinib treatment, we indeed noticed a subset of genes with peak expression during M phase and larger phase delays in D0 than D3 (**Fig. 3.7C**); this included anaphase-promoting complex member *CDC27* (differential delay, dd=0.11 radians), cyclin-dependent kinase inhibitor *CDKN3* (dd=0.10), and centrosome scaffolding factor *ODF2* (dd=0.09) (**Fig. S3.6G**). A decreased cell cycle speed specifically during M phase is consistent with the expected effect of erlotinib, an EGF-blocker inhibiting progression to G1 (Thomas et al., 2007). The result also aligns with evidence that a complete arrest should not be observed for the PC9 cell line, which has been reported to have some resistance to a complete blockade (Lee et al., 2021; Sutter et al., 2006; Ullrich et al., 2008).

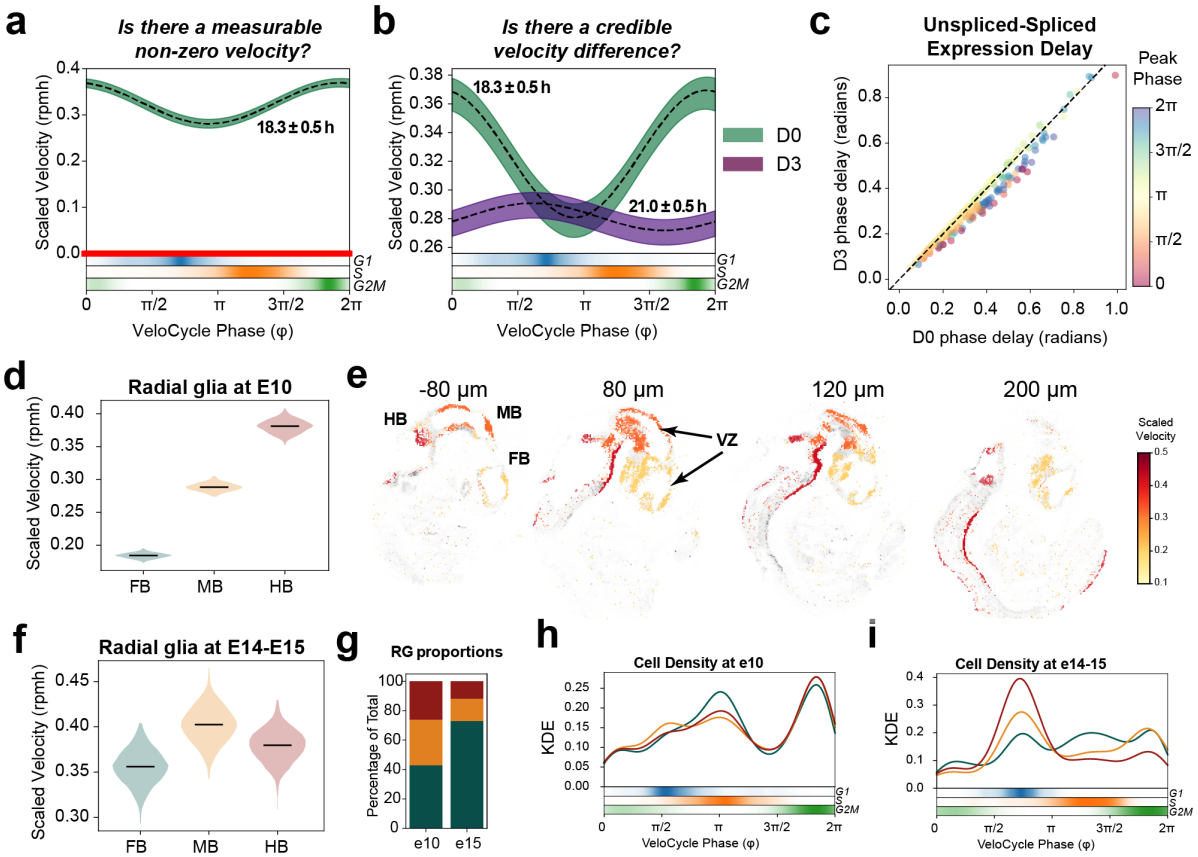


Figure 3.7. *VeloCycle* statistical inference on lung adenocarcinoma and neural progenitors. (a) Posterior estimate plot of scaled velocity in the PC9 lung adenocarcinoma cell line (D0: 9,927 cells) compared to a zero-velocity control (in red) (Aissa et al., 2021). **(b)** Posterior estimate plot of scaled velocity before (D0) and after (D3: 3,943 cells) PC9 treatment with erlotinib. Dashed lines indicate the mean of 500 posterior predictions; green (D0) and purple (D3) bars represent credibility intervals (5th-95th percentile). Areas in which the intervals do not overlap indicate statistically significant velocity differences. Bottom: cell cycle categorical (G1, S, G2/M) phase assignment probabilities. **(c)** Scatter plot of the mean unspliced-spliced expression delay for 273 genes between D0 and D3 samples. Gene dots are colored by peak expression phase. **(d)** Violin plots of scaled velocity estimates obtained for mouse forebrain (FB: 3,293 cells), midbrain (MB: 2,388 cells), and hindbrain (HB: 2,012 cells) radial glial (RG) progenitors at developmental stage E10 (La Manno et al., 2021). Black dashed lines indicate the mean of 500 posterior predictions. **(e)** Spatial projection of single-cell clusters onto four sections of a reference E11 mouse embryo profiled with spatial transcriptomics (hybridization-based in situ sequencing; HybISS), colored by scaled velocity estimates obtained with *VeloCycle*. Regional domains (FB, MB, HB) and the ventricular zone (VZ) are labeled accordingly. Mapping of single cells to spatial data was achieved using the BoneFight algorithm (La Manno et al., 2021). **(f)** Violin plots of scaled velocity estimates for similar populations as on left (FB: 2,460 cells; MB: 307 cells; HB: 176 cells) at developmental stages E14/E15. **(g)** Bar plot of regional proportions of radial glia progenitor cells analyzed at early (E10) and late (E14/E15) time points. **(h)** Kernel density estimation plots of cell distributions along the cell cycle manifold at E10, colored by regional identity. **(i)** Kernel density estimation plots of cell distributions along the cell cycle manifold at E14/E15.

3.3.8. Cell cycle speed in radial glial progenitors varies along a spatio-temporal axis in mouse development

Regulation of proliferation rate as well as of symmetric and asymmetric divisions of radial glia cells (RG) in the ventricular-zone plays a critical role in controlled developmental timing along an anterior-posterior axis of the brain (Beattie & Hippenmeyer, 2017). To elucidate whether there are differences in cell cycle speed among progenitors populating different spatial regions during mouse neurodevelopment, we performed *VeloCycle* estimation on forebrain (FB), midbrain (MB), and hindbrain (HB) RG at the embryonic day 10 (E10) stage (La Manno et al., 2021). Cell cycle speed varied along the forebrain-midbrain-hindbrain axis, with progenitors dividing more quickly posteriorly (HB) than anteriorly (FB) (**Fig. 3.7D**). A finer visualization of this gradient was allowed by computationally mapping the cell cycle speed inferred in these cells to the corresponding locations using in situ hybridization spatial transcriptomics (HybISS) data and the BoneFight algorithm (La Manno et al., 2021) (**Methods 3.5.6.5**). We observed rapidly dividing RG localized close to the ventricular zones, highlighting that cell proliferation takes place along the ventricular zone and suggesting that different segments of the zone proliferate at different rates (**Fig. 3.7E**) (Alieh et al., 2023). Conversely, at E14 and E15 time points, RG from all three brain regions stabilized at a similar proliferation speed, with no credible velocity difference (**Fig. 3.7F**). At these later time points, the majority of RG in the midbrain and hindbrain regions had accumulated in a non-proliferative state; the majority of RG cells present were from the forebrain, which more slowly developed at E10 (**Fig. 3.7G-I**). These results align with recent studies showing that hindbrain specifies into non-proliferating, differentiated cell types more quickly; an increased proliferative capacity is thus likely required in the earlier stages of development (Braun et al., 2023; Di Bella et al., 2021; Ohnuma & Harris, 2003). Furthermore, the later slowdown is expected and in line with what has been reported in EdU tracking studies (Arai et al., 2011; Harris et al., 2018).

3.3.9. Transfer learning of manifold parameters enables discovery of velocity alterations in genome-wide perturbation screens

Previous frameworks for RNA velocity have offered restricted applicability to samples containing few cells. With recent single-cell technologies designed to screen the effects of hundreds of small genetic, environmental, or drug perturbations, there is a growing need to assess changes in cell dynamics under circumstances with limited data (Brunello, 2022; Dixit et al., 2016; Peidli et al., 2023).

VeloCycle, with its manifold-constrained velocity estimates, can explore RNA velocity in such contexts: by transferring *manifold-learning* from a large dataset onto smaller datasets,

one can perform velocity inference using limited cells or using cells representing only a portion of the phase space (**Methods 3.5.6.6**). To demonstrate this, we studied a large-scale, genome-wide Perturb-seq dataset where hundreds of individual gene knockouts were introduced into the RPE1 cell line via a targeted, pooled CRISPR library, followed by scRNA-seq after seven days in culture (Replogle et al., 2022). First, we ran *VeloCycle* on non-targeted control cells (NT) and a pooled group of gene knockout conditions corresponding to well-characterized marker genes for the cell cycle (CC-KO). The cell cycle period was 25.6 ± 1.3 hours for NT and 30.9 ± 1.3 hours for CC-KO (**Fig. 3.8A**). Similar results were obtained when using a large set of cell cycle genes (n=426) compared to a smaller gene set (n=120) (**Fig. S3.7A-E**). When CC-KO conditions were stratified by genes typically considered S and G2/M markers, we observed an accumulation of cells in the G1 phase space compared to NT cells (**Fig. 3.8B, S3.7F**). This suggests the loss of function for some individual cell cycle related genes disrupts cell cycle progression, either by slowing down the proliferation rate in certain phases, or by halting progression altogether ahead of specific entry checkpoints.

To scrutinize the effect of individual gene knockout conditions on cell cycle speed, we employed a transfer learning approach in which we conditioned our manifold-learning on gene harmonics previously inferred from the NT and CC-KO data subsets, assigning phases to a significantly larger population of 167,119 cells and 986 individual knockout conditions, some with as few as 75 cells (**Fig. 3.8C**). Consistent with coarser stratifications of the data, we observed a significant decrease in cell cycle speed in individual cell cycle related-gene knockout conditions compared to both non-targeting control cells and cells with gene knockdown unaffiliated with the cell cycle (**Fig. 3.8D**). Several of the most impaired cell cycle speeds were found in knockouts of highly characterized genes involved in DNA replication (*MCM3Δ* and *MCM6Δ*) and translation initiation (*E1F3BΔ*, *EIF2B3Δ*, *EIF3CLΔ*) (**Fig. 3.8D-E**). Curiously, knockout conditions for several splicing and mRNA processing genes either significantly decreased or increased the estimated cell cycle speed, including *DBR1Δ*, an intron-lariat splicing factor (11.7-fold decrease compared to NT condition), *PRPF3Δ* (1.2-fold increase), and *PRPF31Δ* (1.3-fold increase) (**Fig. 3.8D-E**). Given the dependence of RNA velocity estimation on the governing differential equations of the RNA metabolic life cycle, this result indicated that biological disruptions affiliated to RNA metabolism undermine the biophysical parameterization of the velocity framework. Moreover, the number of cells present in the dataset per condition had a direct influence on the velocity estimate posterior uncertainty, suggesting that more cells, and thereby less aggregated sparsity for a condition, increased the confidence of the *VeloCycle* model in the obtained velocity estimate **(Fig. 3.8F-G**). Ultimately, these analyses demonstrate that velocity can be applied, with transfer learning

approaches, in large-scale perturbation contexts as a metric to assess the impact of gene knockouts on the dynamics of a biological process.

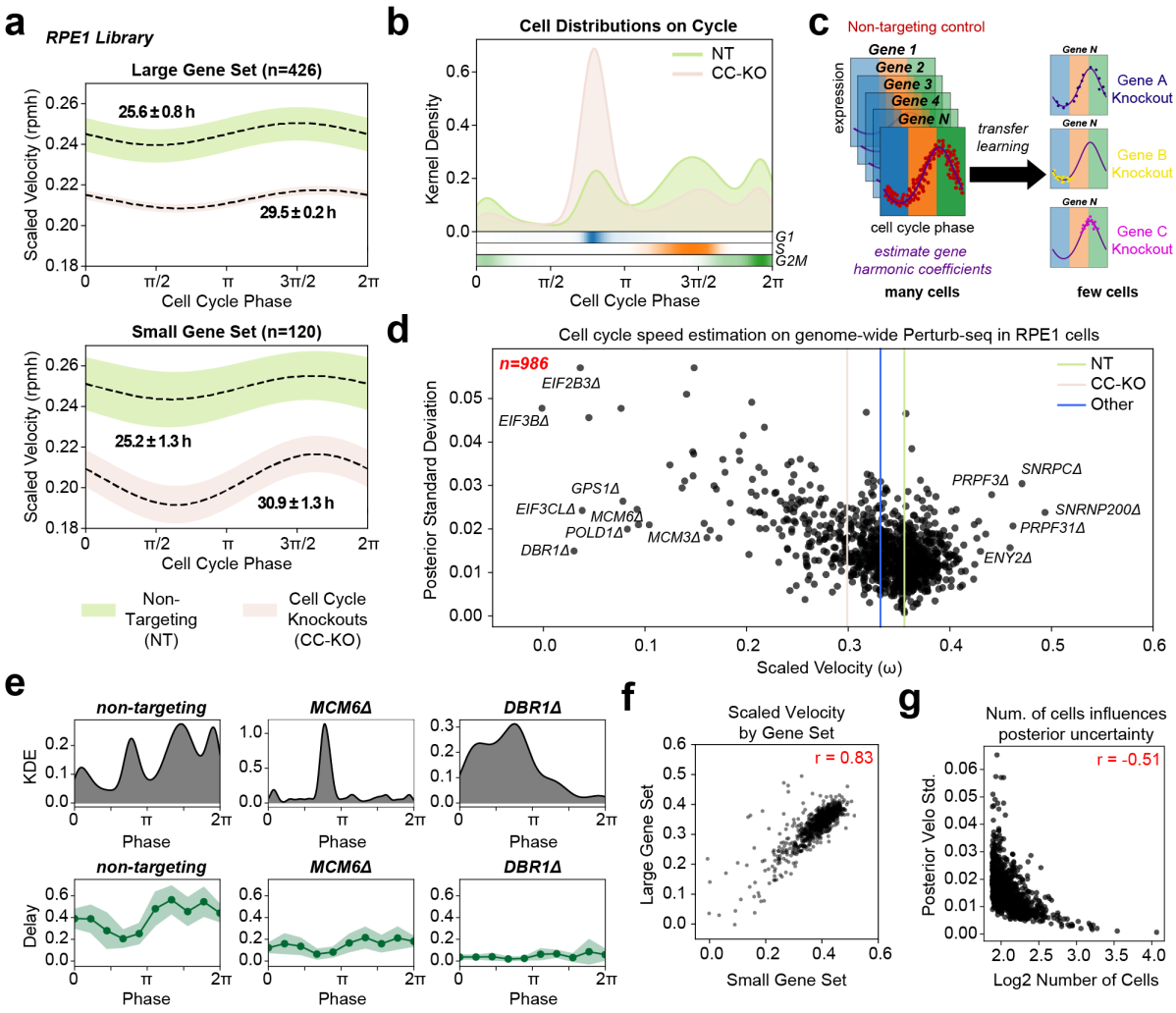


Figure 3.8. Transfer learning of manifold parameters to study effects of genome-wide knockouts on cell cycle velocity. (a) Posterior estimate plot of cell cycle speed for RPE1 cells 7 days after CRISPR-induced single-gene knockdowns with Perturb-seq, stratified by non-targeting controls (green; 11,485 cells) and cell cycle knockout (beige; 6,275 cells) conditions (Replogle et al., 2022). Manifold-learning was performed using either a large (top; n=426) or small (bottom; n=120) gene set. Black dashed lines represent the mean estimate of 500 posterior predictions. **(b)** Kernel density plot of continuous cell cycle phase distributions for non-targeting (NT) and cell cycle knockout (CC-KO) samples from (a). Heatmap bar plot (bottom) is of cell density by categorical phase assignment. **(c)** Schematic of the employed transfer learning approach. Gene harmonic coefficients are obtained on NT controls (with many cells) using the manifold-learning model, and are then applied to assign phases to cells with one of hundreds of gene knockout conditions (each with a few cells) potentially distributed unequally on the manifold. **(d)** Scatter plot of cell cycle *velocity-learning* estimates and posterior standard divisions for 986 individual cell cycle knockout (Δ) conditions in 167,119 RPE1 cells. Vertical lines correspond to the mean velocity estimates for non-targeting (green), cell cycle marker gene knockouts (tan) and other gene knockouts (blue). **(e)** Top: kernel density estimate (KDE) plots of cell distribution for the non-targeting, *MCM6Δ*, and *DBR1Δ* conditions. Bottom: binned unspliced-spliced expression delay (Delay) for the same conditions. Genes were binned by position of peak expression

into 10 groups along the cell cycle to obtain an average delay. The dark green line represents the mean delay; the light green line represents the standard deviation. **(f)** Scatter plot of scaled cell cycle velocity estimates obtained for 986 conditions in (d) using small and large gene sets. **(g)** Scatter plot of total number of cells per condition and posterior velocity standard deviation for 986 conditions in (d). Pearson's correlation coefficient is indicated in red in (f) and (g).

3.4. Discussion

In this work, we address several limitations of current RNA velocity methods by designing a framework that unifies manifold and velocity inference into a single probabilistic generative model. We propose an explicit parametrization of RNA velocity as a vector field defined on the manifold coordinates, implemented and thoroughly tested for one dimensional periodic manifolds and the cell cycle (**Methods 3.5.1-3.5.4**). RNA velocity has been previously applied to illustrate cell cycle progression, yet in ways that required several heuristics and with exclusive exploratory value, as no conclusion could be made from the inferential procedures (Bastidas-Ponce et al., 2019; Gorin et al., 2022; Lo Giudice et al., 2019; Schwabe et al., 2020).

VeloCycle uses variational inference to learn a Bayesian model that operates on raw data with appropriate noise models, instead of heuristic nearest-neighbor smoothing (**Fig. 3.1, Methods 3.5.4**). *VeloCycle* returns uncertainty estimates, enabling direct evaluation of the confidence about the estimation results and cell cycle speed comparisons between samples. These capabilities are relevant in different biological settings, such as in cancer biology, where alterations to the cell cycle progression need to be scrutinized using snapshot single-cell data (**Fig. 3.7**). Therefore, *VeloCycle* could yield new biological insight into disease progression, for example by characterizing differences in proliferation rates between tumors across microenvironments or patients.

Uncertainty measurements are central to statistical evaluation of RNA velocity. The first methods to introduce Bayesian variational inference for RNA velocity modeling, VeloVAE (Gu et al., 2022), VeloVI (Gayoso et al., 2023), and Pyro-Velocity (Qin et al., 2022), simplify the variational distribution in ways that limit usefulness of the estimated joint posterior, particularly given an unscaled gene-wise velocity parametrization. More generally, models with a high number of degrees of freedom and the assumption of independence risk overfitting noise and overestimating confidence in the velocity (Aivazidis et al., 2023; Gorin et al., 2022). In this study, we control for this risk by constraining all spliced-unspliced fits under a single velocity function, and we structure our model to scrutinize dependencies between the cell cycle velocity and kinetic parameters from the full posterior distribution (by MCMC and a LRMN variational distribution) (**Fig. 3.5; Methods 3.5.3.2**). Moreover, we exploit the fact that this parameterization is explicit in the velocity to perform direct inference on a single latent variable.

While our model for RNA velocity estimation offers clear benefits, there remain open avenues for further development. First, while our mathematical framework is amenable to multidimensional formulations and various topologies, the current work focuses on the case of one-dimensional periodic manifolds. Thus, extensions of *VeloCycle* into higher-dimensional latent spaces can be naturally pursued, although significant efforts will be required to find appropriate parametrization of more complex manifolds. Second, the issue of defining dimensionality intersects with that of gene selection; different subspaces defined by unique genes expose distinct manifolds traversed by varying fractions of cells (Gorin et al., 2022). Methods developed with this problem in mind have been recently proposed (Sheng et al., 2023), and with appropriate modifications, these could be integrated into RNA velocity estimation methods to automate topology and gene set selection. In this direction, frameworks that consider multiple manifolds with varying topologies, spanned by cells in different subspaces, while also assigning specific cells and genes to these features, will notably enhance the general applicability and utility of manifold-consistent RNA velocity estimation. Third, our model assumes a constant gene-specific splicing and degradation rate; in fact, for some genes, such rates likely change in different phases of the cell cycle (Battich et al., 2020; Schwabe et al., 2020). A future extension to *VeloCycle* for which the kinetic parameters are defined by a parameterizable function, could address this limitation. Yet, maintaining the model well-conditioned in these settings might be non-trivial.

Widely used standard analysis pipelines use a small group of marker genes to attribute a categorical phase assignment to single cells, even though cell cycle progression is a continuous process (Satija et al., 2015; Wolf et al., 2018). Recent methods to infer continuous phase assignment represent a significant improvement over scoring-based approaches (Auerbach et al., 2022; Liu et al., 2022; Ranek et al., 2022; Zheng et al., 2022). The *manifold-learning* of *VeloCycle* makes progress along this direction, also inferring individual gene periodicity patterns, providing posterior uncertainty and obtaining results that compare favorably with other methods (Riba et al., 2022) (**Fig. 3.3**). Importantly, the *manifold-learning* step is flexible and facilitates transfer learning: the geometry of the manifold can be estimated on a larger or higher quality dataset and serve as a prior for a smaller dataset. This enhances the robustness and applicability of velocity-learning across diverse experimental conditions. We employ these transfer learning capabilities on a Perturb-seq dataset, demonstrating that RNA velocity can be used as a readout of the context of a high-content screen (**Fig. 3.8**). This is particularly relevant given the increased use of barcoding strategies for single-cell level screening. We expect future applications of such models in the context of drug screening and evaluation of genetic changes on heterogeneous pools of cells.

A way to validate the overall consistency of a RNA velocity vector field has been to correlate an heuristically estimated transition probability between populations with prior knowledge on their lineage relationships; however, this is correlative and indirect (Lange et al., 2022). Here, we instead compare directly estimates with the real velocity of the process. By specifically biologically-reasoned priors, velocities obtained with *VeloCycle* can be directly interpreted as the proliferation speed, which can vary in different tissue locations, at different moments of development, or as a result of perturbations to the core gene regulatory network (La Manno et al., 2021; Replogle et al., 2022). This study empirically validated obtained RNA velocities, juxtaposing *VeloCycle* speed estimates with proliferation times obtained by live-cell microscopy imaging and cumulative EdU labeling (**Fig. 3.6**).

Ultimately, our framework represents an advancement in the rigor of dynamical estimations from single-cell data. The promising outcomes of tailoring RNA velocity to single processes advocates for the development of new models that dissect the high-dimensionality of single-cell data into individual biological axes with corresponding and interpretable RNA velocity fields.

3.5. Methods

3.5.1. Model specifications for manifold-constrained RNA velocity

Gene expression measurements as obtained by scRNA-seq provide a high dimensional snapshot of a cell's state, with typically $n \simeq 10^4$ genes being expressed in a cell, of which several thousands are experimentally detected per cell by a nonzero read count. Here, we use the notation $Y_c = (U_c, S_c)$ for the measurements, with U_c for the unspliced and S_c for the spliced RNA levels (counts), with $S_c, U_c \in \mathbb{N}^n$.

3.5.1.1. The manifold

Many biological processes of interest, such as the cell cycle or a differentiation event, unfold on low-dimensional manifolds $\mathcal{M}$. Here, we will consider a parametric representation for $\mathcal{M}: x \mapsto s(x) \in \mathcal{M}$ where x is latent coordinate for each cell c. Moreover, we will choose the manifold topology based on the biological structure of the problem. For example, given a periodic process such as the cell cycle, we will take $x \in S_1$, where S_1 represents the unit circle. Typically, the manifold dimension $m \ll n$ will be small, and in the case of the cell cycle $m = 1$. As we discuss later, we will learn the function $s(x)$ from the data (which we will refer to as the *manifold-learning* procedure).

3.5.1.2. Measurements and noise model

Measurements for each cell c will be linked to the corresponding locations on $\mathcal{M}$ via realistic noise models. In the case of scRNA-seq, relevant noise models consist of negative binomial (NB) distributions, so that $Y_{gc} \sim \mathrm{NB}\big(\mathrm{y_g}(\mathrm{x_c}), \alpha_\mathrm{g}\big)$, with $y_g(x_c) = E\big[Y_{gc}\big] = \big(s_g(x_c), u_g(x_c)\big)$ and $\alpha_g = \big(\alpha_g^s, \alpha_g^u\big)$. Note that we are assuming for simplicity that α_g is independent of x (but this can be relaxed at the expense of an increased number of parameters). This allows us to formulate a likelihood model for the data and approach inference using Bayesian or variational inference.

3.5.1.3. RNA velocity and chemical kinetics

In the high-dimensional gene expression space, we expect a rate equation describing the RNA velocity $\frac{d\tilde{s}}{dt}$ depending both the expectation of spliced and unspliced RNA counts:

$$\frac{d\tilde{s}_g}{dt} = F\big(\tilde{s}_g, \tilde{u}_g\big) = \beta_g \tilde{u}_g - \gamma_g \tilde{s}_g \qquad (\mathbf{1})$$

with time-dependent locations $\tilde{s}_g(t)$ and $\tilde{u}_g(t)$ and gene-dependent RNA splicing and degradation rates β_g and γ_g. Note that here we do not include a corresponding equation for $\frac{d\tilde{u}}{dt}$ as it will not be needed for the application to the cell cycle. Also, F is not explicitly time-dependent and the rates are taken as constants (which could however be relaxed, see below).

3.5.1.4. Latent-space dynamics

The key assumption in our approach is that there exists an autonomous (and here deterministic) equation for the dynamics of $x(t)$:

$$\frac{dx}{dt} = V(x) \qquad (\mathbf{2})$$

which provides a low-dimensional approximation of the full dynamics (Eq. 1), and that $\tilde{s}(t), \tilde{u}(t)$ are time-dependent through $x(t)$:

$$\begin{aligned} \tilde{s}(t) &= s(x(t)) \\ \tilde{u}(t) &= u(x(t)) \end{aligned} \qquad (\mathbf{3})$$

$V(x)$ is the vector field describing the dynamics in the low dimensional latent space.

3.5.1.5. Manifold-constrained RNA velocity

We can now link Eqs. (1), (2) and (3) to obtain

$$\frac{ds_g(x(t))}{dt} = \big(\nabla_x s_g\big) \cdot V\big(x(t)\big) = \beta_g u_g\big(x(t)\big) - \gamma_g s_g\big(x(t)\big)\ \forall g \qquad (\mathbf{4})$$

where we have introduced the gene index g for clarity and applied the chain rule. β_g and γ_g are the gene-specific splicing and degradation rates.

Eq. (4) provides the basis of our approach as it connects the topology of the low dimensional manifold on the left-hand side with the biology on the right-hand side. Of note, the parameters governing gene dynamics (β, γ) could in principle depend on x as well.

3.5.1.6. Geometric interpretation

By construction, we see that the RNA velocity vector $\frac{ds(x(t))}{dt}$ lies in the tangent space of $\mathcal{M}$ at every point of a trajectory $s_g(x(t))$. Indeed $\nabla_x s$ forms an m-dimensional basis of the tangent space at each point and $V(x(t))$ forms the components of the velocity vector in that basis.

3.5.1.7. $u(x)$ and inference

Eq. (4) can also be viewed as specifying $u(x)$ given $s(x), V(x)$, and the parameters β and γ. This will become central in the implementation. In essence, the optimization algorithm to identify $V(x)$ and γ and β coefficients (or functions if we would allow $\gamma = \gamma(x)$, etc) such that the predicted RNA velocity $\frac{ds(x(t))}{dt}$ (which lies in tangent space over the entire manifold $\mathcal{M}$) is closest to that implied by chemical kinetics and the data $Y_c = (U_c, S_c)$.

3.5.1.8. Duration of biological processes

A benefit of this formulation is that it becomes accessible to estimate the actual duration of biological processes from the trajectories and $V(x)$:

$$\Delta t_{s_0,s_1} = \int_{\Gamma_{s_0}^{s_1}} \frac{1}{\dot{s}} ds = \int_{\Gamma_{x_0}^{x_1}} \frac{1}{v(x)} dx = \Delta t_{x_0,x_1} \qquad (\mathbf{5})$$

where $\Gamma_{x_0}^{x_1}$ is the trajectory $x(t)$ that connects the two points x_0 and x_1, and where we have used the change of trajectory variable $s(x)$. For example, we will be able to estimate cell cycle periods. Moreover, this estimate is by construction independent of the parametrization of the low dimensional manifold.

3.5.2. Manifolds with S^1 topology: the cell cycle

Here, we assume that $\mathcal{M}$ is topologically a circle and therefore we write the coordinate x as $\varphi \in S^1$. The equation of the dynamics (Eq. 4) becomes

$$\dot{s}_g = \partial_\varphi s_g(\varphi)\omega(\varphi) = \beta_g u_g - \gamma_g s_g \qquad (\mathbf{6})$$

$$E[S_{gc}] = s_g(\varphi_c) = \exp\left(\sum_f \nu_{gf}\zeta_f(\varphi_c)\right) \quad (7)$$

where we assume that β_g and the γ_g are constant along the cell cycle. Of note is that the values of those parameters are constrained by the biology (see section 4 below), which we will enforce through appropriate priors.

S^1 is convenient since it allows use of Fourier series to parameterize the various functions: $s(\varphi), u(\varphi), \omega(\varphi)$. Typical cell cycle genes exhibit profiles that can be described by only few harmonics; thus, we will consider up to k Fourier components in our expansion (in practice we will by default use one harmonic). Moreover, since $s(\varphi)$ is positive, we will use the notation

$$\log\left(s_g(\varphi_c)\right) = \sum_f \nu_{gf}\zeta_f(\varphi_c) \quad (8)$$

with

$$\nu_g = \begin{pmatrix} a_g^0 \\ a_g^1 \\ b_g^1 \\ \vdots \\ a_g^k \\ b_g^k \end{pmatrix} \quad \zeta(\varphi) = \begin{pmatrix} 1 \\ \cos(\varphi) \\ \sin(\varphi) \\ \vdots \\ \cos(k\varphi) \\ \sin(k\varphi) \end{pmatrix} \quad (9)$$

Here ν_g is the vector of gene Fourier parameters written with real numbers.

Using the chain rule, we obtain $u(\varphi)$:

$$\partial_t s_g(\varphi) = \omega(\varphi) s_g(\varphi) \sum_f \nu_{gf}\, \partial_\varphi \zeta_f(\varphi) \quad (10)$$

which leads to

$$\log\left(u_g(\varphi)\right) = -\log(\beta_g) + \log\left(\omega(\varphi)\sum_f \nu_{gf}\, \partial_\varphi \zeta_f(\varphi) + \gamma_g\right) + \log\left(s_g(\varphi)\right) \;\; \forall g \quad (11)$$

$$E[U_{gc}] = u_g(\varphi) = \frac{s_g(\varphi)}{\beta_g}\left(\omega(\varphi)\sum_f \nu_{gf}\, \partial_\varphi \zeta_f(\varphi) + \gamma_g\right) (12)$$

For $\omega(\varphi)$ we will also be using a Fourier series, limiting ourselves to either constant ω or $\omega(\varphi)$ functions with one harmonic.

3.5.2.1. Likelihoods

As explained above with the expressions for $u(\varphi)$ and $s(\varphi)$, we can calculate a likelihood for the count data over all cells $\{Y_c\} = \{(U_c, S_c)\}$. To simplify the implementation, we approximate the full joint likelihood for $\{(U_c, S_c)\}$ as a product of two factors:

$$P(\{(S_c, U_c)\} \mid \theta) = \prod_{cg} P\left(S_{cg}, U_{cg} \mid \omega(\varphi), \varphi_c, \nu_g, \beta_g, \gamma_g, \alpha_g\right) \text{ with}$$

$$P\left(S_{cg}, U_{cg} \mid \theta\right) = P_s\left(S_{cg} \mid \nu_g, \alpha_g^s, \varphi_c\right) \times P_u\left(U_{cg} \mid \omega(\varphi), \beta_g, \gamma_g, \nu_g, \varphi_c, \alpha_g^u\right) \quad (13)$$

$$P_s\left(S_{cg} \mid \cdots\right) = \mathrm{NB}\left(s_g(\varphi_c) = F\left[\nu_g, \varphi_c\right], \alpha_g^s\right), \quad (14)$$

$$P_u\left(U_{cg} \mid \cdots\right) = \mathrm{NB}\left(u_g(\varphi_c) = G\left[\omega(\varphi_c), \beta_g, \gamma_g, \nu_g, \varphi_c\right], \alpha_g^u\right) \quad (15)$$

where θ is a generic notation for parameters, and $F[\ldots], G[\ldots]$ show the dependencies of s_g, u_g on the other quantities.

We combine these likelihoods with a set of priors into a full Bayesian model (see below) to estimate the joint posterior of θ. As indicated above, in our current implementation we simplify the problem by taking two steps: first, we optimize P_s to estimate the cell phases $\{\varphi_c\}$ and Fourier coefficients $\{\nu_g\}$. We call this step the *manifold-learning* procedure. The second step optimizes P_u and is called *velocity-learning*, using the posterior expectations for $\left(\{\varphi_c\}, \{\nu_g\}, \{\alpha_g^s\}\right)$ obtained during *manifold-learning* to estimate the remaining quantities $\left(\omega(\varphi), \beta_g, \gamma_g, \alpha_g^u\right)$.

3.5.3. Bayesian model formulation for *VeloCycle*

Our model includes a mix of biologically defined priors with Empirical Bayes-style priors determined from the data. Our goal will be to estimate an approximation of the joint posterior probability distribution, based on the above expression of the likelihoods:

$$P(\theta \mid \{S_c, U_c\}) = \frac{P(\{S_c, U_c\} \mid \theta) P(\theta)}{P(\{S_c, U_c\})} = \frac{\prod_{cg} P\left(S_{cg} \mid \theta\right) P\left(U_{cg} \mid \theta\right) P(\theta)}{\int \prod_{cg} P\left(S_{cg} \mid \theta\right) P\left(U_{cg} \mid \theta\right) P(\theta) d\theta}$$

We specify the following priors $P(\theta)$.

$$\nu\omega_t \sim \mathcal{N}([0,0,0], [3^2, 0.05^2, 0.05^2])$$
$$\log\left(\gamma_g\right) \sim \mathcal{N}(0, 0.5^2)$$
$$\log\left(\beta_g\right) \sim \mathcal{N}(2, 3^2)$$
$$\alpha_g \sim \mathrm{Gamma}(1.0, 2.0)$$
$$\nu_{gt} \sim \mathcal{N}\left(\mu_{gt}^{\nu}, {\sigma_{gt}^{\nu}}^2\right)$$
$$\varphi x y_c \sim \mathrm{ProjNormal}(\varphi x_c, \varphi y_c)$$

Setting by empirical Bayes the following parameters:

$$\mu_{gt}^{\nu} = \left[\log\left(\mathrm{mean}_c\left(S_{cg}\right)\right), \mathbf{0}, \mathbf{0}\right]$$

$$\sigma_{gt}^{\nu} = \begin{bmatrix} \frac{1}{2}\mathrm{std}_c\left(S_{cg} + 1\right), \\ \frac{1}{4}\mathrm{std}_c\left(S_{cg} + 1\right), \\ \frac{1}{4}\mathrm{std}_c\left(S_{cg} + 1\right) \end{bmatrix}$$

$$\varphi x_c = \varepsilon \cos(\Phi_c)$$
$$\varphi y_c = \varepsilon \sin(\Phi_c)$$

where Φ_c is obtained from the two first principal components (w_{1c}, w_{2c}) renormalized between $[-0.5, 0.5]$ and computing $\Phi_c = \tan^{-1}(w_{2c}, w_{1c})$. Rotational invariance (e.g., arbitrariness of the first cell $c0$ so that $\Phi_{c0} = 0$) is obtained by finding the global phase shift maximizing $\text{corr}\left(\Phi_c, \sum_g S_{cg}\right)$. The concentration parameter of the projected normal ε is set to 5 by default, but can be adjusted depending on the overall confidence in the data quality.

3.5.3.1. Variational Distribution - SVI

The variational distribution we use in the base model is mean-field, with marginals of either Normal or Dirac Delta distributed. Specifically

$$P\left(\{\nu\omega_t\}, \{\varphi_c\}, \{\nu_{gt}\}, \{\beta_g\}, \{\gamma_g\}, \{\alpha_g\}\right) = \prod_c \prod_g \prod_t P(\nu\omega_t) P(\varphi_c) P\left(\nu_{gt}\right) P\left(\beta_g\right) P\left(\gamma_g\right) P\left(\alpha_g\right)$$

The variational distribution is parametrized as follows (^ indicates the parameters):

$$\begin{aligned}
&P(\nu\omega_t) = \mathcal{N}(\widehat{\mu\nu\omega}_t, \widehat{\sigma\nu\omega}_t^2) \\
&P\left(\nu_{gt}\right) = \mathcal{N}\left(\widehat{\mu_{gt}^{\nu}}, \widehat{\sigma_{gt}^{\nu}}^2\right) \\
&P\left(\alpha_g\right) = \text{Delta}\left(\widehat{\alpha_g}\right) \\
&P\left(\log\left(\gamma_g\right)\right) = \mathcal{N}\left(\widehat{\mu\log\gamma}_g, \widehat{\sigma\log\gamma}_g^{\,2}\right) \\
&P\left(\log\left(\beta_g\right)\right) = \mathcal{N}\left(\widehat{\mu\log\beta}_g, \widehat{\sigma\log\beta}_g^{\,2}\right) \\
&P(\varphi x y_c) = \mathcal{N}([\widehat{\varphi x}_c, \widehat{\varphi y}_c], [1,1])
\end{aligned}$$

3.5.3.2. Variational Distribution - LRMN

The Low Rank Multivariate Normal (LRMN) model considers a variational distribution parametrized to mimic the correlative structure observed between the joint posteriors sampled by Markov Chain Monte Carlo (MCMC) estimation. Specifically, we allow for a covariance and establish specific conditional relationships between the velocity, or angular speed $\nu\omega_t$, and the kinetic parameters β_g and γ_g. The two main features are: (a) the joint posterior between γ_g and $\nu\omega_t$ is parametrized as a low-rank Multivariate Normal, and (b) the marginal posterior of β_g is expressed as conditioned on γ_g; namely for each gene g, the marginal posterior of β_g , through an explicit parameter $\widehat{\rho_g}$, is allowed to correlate with the correspondent γ_g.The posterior factorizes as follows:

$$P\left(\{\nu\omega_t\}, \{\varphi_c\}, \{\nu_{gt}\}, \{\beta_g\}, \{\gamma_g\}, \{\alpha_g\}\right) =$$

$$= P\left(\{\gamma_g\}, \{\nu\omega_t\}\right) \prod_g P\left(\beta_g \mid \gamma_g\right) P\left(\alpha_g\right) \prod_t P(\nu\omega_t) P\left(\nu_{gt}\right) \prod_c P(\varphi_c)$$

The specific formulation we used is:

$$\mathbf{x} \equiv \left[\log(\gamma_1), \log(\gamma_2), \ldots., \log\left(\gamma_{n_g}\right), \nu\omega_0, \nu\omega_1, \ldots, \nu\omega_{n_{t-1}}\right]$$

$$\boldsymbol{\Sigma} = \hat{\mathbf{F}}\hat{\mathbf{F}}^{\mathrm{T}} + \mathrm{diag}(\hat{\mathbf{d}}) \text{ where } \hat{\mathbf{F}} \in \mathbb{R}^{(n_g+n_t)\times k}, \text{ with } k = 5$$

$$P(\{\log(\gamma_g)\}, \{\nu\omega_t\}) = P(\mathbf{x}) = \mathrm{MultivariateNormal}(\hat{\boldsymbol{m}}, \boldsymbol{\Sigma})$$

$$\mu \log \beta_g \mid \gamma = \mu\widehat{\log \beta_g} + \widehat{\rho_g} \cdot \mu\widehat{\log \beta_g} \cdot \frac{(\log(\gamma_g) - \mu\widehat{\log \gamma_g})}{\sigma\widehat{\log \gamma_g}} \text{ with } \widehat{\rho_g} \in [0,1]$$

$$\sigma \log \beta_g \mid \gamma = \mu\widehat{\log \beta_g}\sqrt{1 - \widehat{\rho_g}^2}$$

$$P\left(\log(\beta_g) \mid \log(\gamma_g)\right) = \mathcal{N}\left(\mu\log\ \beta_g|\gamma, \sigma\log\ \beta_g|\gamma^2\right)$$

$$P(\varphi x y_c) = \mathcal{N}([\widehat{\varphi x}_c, \widehat{\varphi y}_c], [1,1])$$

$$P(\nu_{gt}) = \mathcal{N}\left(\widehat{\mu_{gt}^{\nu}}, \widehat{\sigma_{gt}^{\nu}}^2\right)$$

$$P(\alpha_g) = \mathrm{Delta}(\widehat{\alpha_g})$$

3.5.4. Model implementation

To estimate an approximation of the joint posterior probability distribution for the angular cell cycle speed ($\nu\omega_t$) and the parameters of the S^1 manifold upon which $\nu\omega_t$ unwinds, we formulate a likelihood model for the data that we then solve using variational inference in Pyro (Bingham et al., 2018.). This implementation performs estimation of the model latent variables in two steps: *manifold-learning* and *velocity-learning* (**Table 3.1**).

For *manifold-learning*, we estimate the position of each cell along the circular cell cycle manifold (φ) as well as the Fourier series coefficients for each gene (ν) describing their periodicity. These variables are then used to model the expectation of log spliced counts (ElogS), which are themselves modeled from the real data and a Negative Binomial. We initialize all variables to the mean of the prior, which is determined using either the first two principal components (φ) or the per-gene mean and standard deviations of spliced expression (ν). To allow for differences in average expression levels between different datasets or batches, we also define an offset term ($\Delta\nu$) for the first gene harmonic coefficient.

For *velocity-learning*, we infer the Fourier coefficients of the angular speed ($\nu\omega$) as well as velocity kinetic parameters (γ and β), conditioned on the mean of the posterior estimates for parameters obtained during *manifold-learning*. These variables are used to model the expectation of log unspliced counts (ElogU), which are themselves modeled from the real data and a Negative Binomial. We initialize all variables to the mean of the prior, which is zero for the angular speed (an assumption of zero cell cycle velocity). In order to enforce positive $\left(\omega(\varphi)\sum_f \nu_{gf}\, \partial_\varphi \zeta_f(\varphi) + \gamma_g\right)$ in Eq. (10) during learning, we use a relu function.

Given data, we solve the *VeloCycle* model using stochastic variational inference (SVI) and apply a ClippedAdam optimizer and ELBO loss function, with an evolving learning rate

decaying from 0.03 to 0.005 from the first to last training iteration. Typically, we perform 5,000 training iterations for *manifold-learning* and 10,000 training iterations for *velocity-learning*. However, an option to terminating training early is made available, such that no further iterations are executed if the mean loss during the previous 100 iterations is less than 5 units different from the mean loss during the previous 10 iterations.

When performing Monte Carlo Markov Chain (MCMC), we use a No-U-Turn (NUTS) kernel beginning the mean posterior estimates obtained first with SVI. We typically use one chain, 2,000 warm-up sampling steps, and 500 real sampling steps.

VeloCycle can be run using either a local CPU or GPU in a few minutes, with significantly improved runtime speeds on GPU, particularly when using a large number of cells (>30,000 cells) or genes (>300 genes). Since there are many more parameters along the gene dimension, scaling up the number of genes reduces runtime more quickly than scaling up the number of cells.

3.5.4.1. Biological constraints on parameters

Velocity kinetic parameters β and γ are constrained by the biology. In particular,

$$\gamma_g^{-1} \in [0.5,1.5] \text{ hours}$$
$$T = 2\pi/\omega_o \in [6,50] \text{ hours}$$

Moreover, the priors for the gene harmonic coefficients are determined for each gene based on the mean level of expression and the variance across all the cells in the data. For the velocity harmonic coefficients, we assume as a prior mean no velocity (i.e., 0) with a wide standard deviation (3.0).

All priors can be easily modified using the velocycle.preprocessing suite of functions and provided to a Pyro model object using the metaparameters (mp) term.

3.5.4.2. Approximate point estimate for constant cell cycle velocity

To gain an initial insight into the relationship between cell cycle velocity and the expression profiles of the unspliced (u)/spliced (s) read counts, we used a simplified calculation based on solving the first order differential equation $\frac{d}{dt}s_g(t) = \beta_g u_g - \gamma_g s_g$, where the degradation rate γ_g is a gene-dependent constant. If we assume that $u_g(t)$ follows a periodic function with a single harmonic, i.e. $u_g(t) = u_{0g}\left(1 + \varepsilon\cos\left(\omega t - \varphi_{0g}\right)\right)$, then $s_g(t)$ has the same functional form but with a scaled amplitude and shifted phase, depending on the half-life: $s_g(t) = s_{0g}\left(1 + \varepsilon'\cos\left(\omega t - \varphi_{1g}\right)\right)$, with $\varepsilon' = \varepsilon\cos\left(\Delta\varphi_g\right), \Delta\varphi_g = \left(\varphi_{1g} - \varphi_{0g}\right)$ and $\tan\left(\Delta\varphi_g\right) = \omega\gamma_g^{-1}$. Here, ω represents the cell cycle velocity.

Assuming now that we have multiple conditions (or replicates) c and that the half lives $\tau_g = \gamma_g^{-1}$ are condition-independent, we observe that the relation

$$\delta_{cg} = \tan(\Delta\varphi_{cg}) = \omega_c \tau_g$$

is a rank-1 decomposition of the matrix δ_{cg}, which can be computed using the singular value decomposition (SVD), i.e., $\delta_{cg} = u_c d v_g +$ higher rank terms, using standard notation. This allow us to express the condition-specific cell-cycle velocity ω_c in units of inverse mean half lives (noted $\omega_c^\star$) as

$$\omega_c^\star = u_c d \overline{v_g}$$

where $\overline{v_g}$ stands for the mean over genes. The cycle-cycle period in units of mean half lives is then $T_c^\star = \frac{2\pi}{\omega_c^\star}$.

3.5.4.3. Gene sets and quality control filtering

To select genes for velocity analysis that are expected to behave periodically with the cell cycle, we applied one of three differently-sized, literature-based cycling gene sets: “small” containing 97 genes (Satija et al., 2015), “medium” containing 218 genes (Riba et al., 2022), and “large” containing 1,918 genes (Ontology Consortium et al., 2023). *VeloCycle* uses the “medium” gene set as a default, in order to minimize the influence of noisy or lowly-expressed genes on manifold and velocity estimation; however, we also employed the “large” gene set in contexts where the sequencing depth and dataset quality are particularly high. The command velocycle.utils.get_cycling_gene_set() can be used to access these human and mouse gene sets. Additional gene filtering based on mean detection of spliced and unspliced counts was also performed as described in the sections below.

3.5.4.4. Categorical and continuous cell cycle phase assignment

Categorical cell cycle phase assignment (G1, S, G2/M) was performed using the scanpy function sc.tl.score_genes_cell_cycle(), as previously described (Satija et al., 2015). Continuous cell cycle phase assignment using DeepCycle on both simulated and real datasets was achieved using the velocity information obtained from scvelo.pp.moments (Bergen et al., 2020) and standard parameters described in the original publication (Riba et al., 2022).

3.5.4.5. Inference of the unspliced-spliced delay

To compute the unspliced-spliced delay from the results of *VeloCycle*, we calculated the difference between phases of peak expression of unspliced and spliced UMIs on a per

gene basis (in radians) using the estimated expectations of unspliced (ElogU) and spliced (ElogS) counts.

3.5.4.6. Posterior probability sampling

Unless otherwise stated, the latent variables and associated estimate uncertainties were collected from 500 posterior samples after model training using pyro.infer.predictive, and credibility intervals were measured between the 5th and 95th percentiles.

Estimates for the cell cycle velocity obtained from the velocity function $\omega(\phi)$ were scaled by the mean degradation half-life, i.e., mean(γ_g). To infer the cell cycle period over the entire cell cycle, we sampled from the velocity function on a grid of 20 phases (from 0 to 2π) and took the area under the curve using scipy.integrate.trapz. The posterior mean, 5th percentile, and 95th percentile were then computed using numpy.mean and numpy.percentile. The full uncertainty range of the posterior estimate was computed by taking the difference between the 95th and 5th percentile estimates.

3.5.5. Structured data simulations and sensitivity analyses of *VeloCycle*

To properly validate the performance of *VeloCycle* on datasets with a ground truth for all latent variables of the *manifold-learning* and *velocity-learning* procedures, we employed a new structured simulation approach to preserve relationships among velocity kinetic parameters (splicing rate β, degradation rate γ) and gene harmonics (v_0, $v_{1\sin}$, $v_{1\cos}$). These relationships are expected in real data (La Manno et al., 2018) and are necessary in simulations to avoid improbable scenarios where the ratio of unspliced to spliced counts is unrealistically high or low. We expect that genes containing more velocity information should be those with a larger unspliced-spliced delay and slower splicing and degradation rates; genes with too fast kinetics will provide limited signal in scRNA-seq data. Thus, we formulated a generative *VeloCycle* model that imposes a correlation structure among the gene harmonic and velocity kinetic rate parameters for the sole purpose of sampling simulated data (and not for use during inference itself). We defined correlations as follows: a weak positive correlation among the gene harmonic coefficients (r=0.05; assuming only one sine and cosine term per gene), a moderate positive correlation between the splicing rate and zeroth gene harmonic coefficient v_0 (r=0.30), and a moderate positive correlation between splicing and degradation rates (r=0.30).

Using this correlation matrix, simulated datasets were generated by randomly sampling for a user-defined number of genes and cells, from a pyro.dist.MultivariateNormal. These variables, along with a user-defined ground-truth cell cycle speed ($v\omega$) and a cell-

specific phase (ϕ) sampled from a random uniform distribution between 0 and 2π, were plugged into the velocity equations to compute an expectation for unspliced (ElogU) and spliced (ElogS). Finally, raw data (S and U) was sampled from a pyro.dist.GammaPoisson using the expectations and a noise parameter (shape_inv) sampled from pyro.dist.Gamma. All simulated data generated for this study are available on Zenodo (see **Data Availability 3.6.1**). Additional datasets can be simulated using the velocycle.utils.simulate_data function. Evaluation of the *manifold-learning* step was performed using 20 datasets, each containing 3,000 cells and 300 genes, independently simulated with a ground truth velocity of 0.4. The same datasets were also used for validation of the *velocity-learning* step. To perform sensitivity analysis on the number of cells and genes, four independently simulated datasets containing 10,000 cells and 1,000 genes were generated; data subsets were used to test the model's performance on varying numbers of cells (from 100 to 5,000 for *manifold-learning* and from 50 to 10,000 for *velocity-learning*) and genes (from 100 to 1,000 for *manifold-learning* and from 50 to 1,000 for *velocity-learning*). To assess *velocity-learning* performance on datasets with different ground truth velocities, we simulated four datasets with shared kinetics and gene harmonic parameters, but one of 16 different ground truth velocities from 0.0 to 1.5.

Circular correlations between estimated and simulated ground truth variables were computed using velocycle.utils.circular_corrcoef, which converts the input data into unit circle coordinates and computes a correlation by finding the mean of the product of estimated values and the complex conjugate of the ground truth values. To compare phases obtained with *VeloCycle* to those from DeepCycle, the same simulated datasets were used to compute velocity moments with sc.pp.moments followed by running DeepCycle with default parameters described in the original publication (Riba et al., 2022).

3.5.6. *VeloCycle* estimation across multiple standard scRNA-seq datasets

In this work, we performed manifold geometry and cell cycle velocity estimation with *VeloCycle* on a number of published datasets from different technologies, species, and sampling contexts. For all datasets, the original raw data were re-processed using *velocycle* (La Manno et al., 2018) to obtain spliced and unspliced count matrices. A general procedure for running *VeloCycle* on these types of scRNA-seq data has been described above and is supported by tutorials on the corresponding GitHub page for these works. Here, we provide a summary of any specific filtering criteria and parameters used on a dataset-dependent basis.

3.5.6.1. FACS-sorted mouse embryonic stem cells (Buettner et al., 2015)

VeloCycle estimation of cell cycle phases and gene harmonics was performed on 279 single cells from a culture of Smart-seq2 mouse embryonic stem cells (mESC) using the standard parameters. Genes used in *manifold-learning* were those from the "large" gene set (Gene Ontology; 1,918 genes) available in velocycle.utils, after filtering out genes with <=0.5 mean unspliced counts per cell or with <=2 mean spliced counts per cell (1,358 genes remaining). *Manifold-learning* was performed using 3,000 training steps.

To evaluate the predictive capacity of categorical cell cycle phase (G1, S, or G2/M) using the *VeloCycle* phases, a DecisionTreeClassifier from sklearn.tree was trained with 65% of cells, reserving 35% of cells as a test set and for calculation of a confusion matrix. To compare with a model using the total gene expression matrix to predict categorical cell cycle phases, the linear LogisticRegressionCV model from sklearn.linear_model was trained (cv=5) using the same train-test cell split as with the decision tree.

3.5.6.2. Mouse embryonic stem cells and human fibroblasts (Riba et al., 2022)

VeloCycle was run separately on 5,191 single cells from a culture of mouse embryonic stem cells (mESC) and on 2,557 single cells from a culture of human fibroblasts using the standard parameters. Non-cycling cells were filtered out prior to analysis according to the author's annotations. Genes used in *manifold-learning* were those from the "medium" gene set (DeepCycle; 218 genes) available in velocycle.utils, after filtering out genes with <=0.1 mean unspliced UMIs per cell or with <=0.3 mean spliced UMIs per cell (189 genes and 160 genes remaining for mESC and fibroblasts, respectively). *Manifold-learning* was performed using 5,000 training steps, and *velocity-learning* was performed using the "normal" guide and the constant-velocity model for 10,000 training steps. Comparisons to DeepCycle phases were made using the published estimates described for these exact datasets in the original study.

3.5.6.3. Human dermal fibroblasts (Capolupo, Khven, et al., 2022)

VeloCycle was run on 1,222 single cells from a culture of untreated dermal human fibroblasts (dHFs) using the standard parameters; non-cycling cells were excluded using the author's annotations. Genes used in *manifold-learning* were those from the "large" gene set (Gene Ontology; 1,918 genes) available in velocycle.utils, after filtering out genes with <=0.1 mean unspliced UMIs per cell or with <=0.3 mean spliced UMIs per cell (876 genes remaining). *Manifold-learning* was performed using 5,000 training steps, and *velocity-learning* was performed with both the constant-velocity and periodic-velocity models for 10,000 training steps using the low-rank multivariate normal ("lrmn") guide.

Time-lapse microscopy data, including cell segmentation and tracking, for dHFs was obtained from the originally-published study and is available on the corresponding Zenodo page: 10.5281/zenodo.6245943. A cell was determined to be poorly-tracked and excluded from analysis if it had a measured cell cycle length less than 8 hours or greater than 32 hours.

3.5.6.4. PC9 lung adenocarcinoma cell line (Aissa et al., 2021)

VeloCycle was run jointly on data from PC9 lung adenocarcinoma cell line prior to (D0: 7,927 cells) and after (D3: 3,743 cells) treatment with erlotinib using the standard parameters. Genes used in *manifold-learning* were those from the "large" gene set (Gene Ontology; 1,918 genes) available in velocycle.utils, after filtering out genes with <=0.1 mean unspliced UMIs per cell or with <=0.1 mean spliced UMIs per cell. After an initial *manifold-learning* step, only genes with a Pearson's correlation between the unspliced and spliced counts >=0.8 and a predicted unspliced-spliced delay greater than >=-0.25 were retained. *Manifold-learning* was performed using 5,000 training steps, and *velocity-learning* was performed using the "lrmn" guide and both the constant-velocity and periodic-velocity models for 10,000 training steps.

3.5.6.5. Radial glial progenitors from the developing mouse brain (La Manno et al., 2021)

VeloCycle was run jointly on all radial glia progenitor cells from the E10 time point, stratified by regional identity (forebrain: 3,293 cells; midbrain: 2,388 cells; hindbrain: 2,012 cells) using the standard parameters. Genes used in *manifold-learning* were those from the "large" gene set (Gene Ontology; 1,918 genes) available in velocycle.utils, after filtering out genes with <=0.05 mean unspliced UMIs per cell or with <=0.1 mean spliced UMIs per cell. After an initial *manifold-learning* step, only genes with a Pearson's correlation between the unspliced and spliced counts >=0.8 and a predicted unspliced-spliced delay greater than >=-0.10 were retained. *Manifold-learning* was performed using 5,000 training steps, and *velocity-learning* was performed using the "lrmn" guide and the constant-velocity model for 10,000 training steps.

Similarly, *VeloCycle* was run jointly on all radial glia progenitor cells from the E14 and E15 time points, stratified by regional identity (forebrain: 2,460 cells; midbrain: 307 cells; hindbrain: 176 cells) using the standard parameters. With the same gene filtering steps as with the E10 analysis above, 239 genes were used. *Manifold-learning* was performed using 5,000 training steps, and *velocity-learning* was performed using the "lrmn" guide and the constant-velocity mode for 10,000 training steps.

To spatially visualize *VeloCycle* speed estimates at the E10 time point, we ran the BoneFight algorithm to map scRNA-seq clusters to a corresponding spatial transcriptomics

dataset of hybridization-based in situ sequencing (HybISS) from the same study, then colored the corresponding clusters by their velocity estimate.

3.5.6.6. Genome-wide Perturb-seq RPE1 cells data (Replogle et al., 2022)

To ensure analysis was performed only on RPE1 cells with a complete knockdown of the individual gene target, we filtered out cells containing non-zero unspliced or spliced UMI reads for the targeted gene. *VeloCycle* was run initially on a subset of data in two conditions: (1) the set of control, non-targeting cells (NT: 11,485 cells) and a grouped set of cells where a gene from the "small" cell cycle marker list were targeted for knockdown (CC-KO: 6,275 cells). Genes used in *manifold-learning* were those from the "medium" gene set (DeepCycle; 218 genes) available in velocycle.utils, after filtering out genes with <=0.1 mean unspliced UMIs per cell or with <=0.2 mean spliced UMIs per cell. After an initial *manifold-learning* step, only genes with a Pearson's correlation between the unspliced and spliced counts >=0.7 and a predicted unspliced-spliced delay greater than >=-0.5 were retained (120 genes remaining). *Manifold-learning* was performed using 5,000 training steps, and *velocity-learning* was performed using the "lrmn" guide and the constant-velocity mode for 10,000 training steps.

To perform stratified analysis on a larger batch of gene knockout conditions, we selected all cells from any conditions represented by more than 75 cells, leaving a total of 167,119 cells and 986 conditions. We then conditioned on the gene harmonic coefficients obtained with the coarser analysis using NT and CC-KO cells, and performed *manifold-learning* for 5,000 training steps to estimate cell cycle phases for all cells and conditions. We then performed *velocity-learning* for 10,000 training steps on the entire dataset, estimating an individual constant velocity for each gene knockout condition.

3.5.6.7. RPE1 cells (newly-generated for this study)

To estimate the unspliced-spliced delay and cell cycle velocity between two identical replicates of FUCCI-RPE1 cells (replicate 1: 4,265 cells; replicate 2: 9,994 cells), *manifold-learning* was run on the "medium" gene set available in velocycle.utils, after filtering out genes with <=0.1 mean unspliced UMIs per cell or with <=0.3 mean spliced UMIs per cell (136 genes remaining). *Manifold-learning* was performed using 3,000 training steps, and *velocity-learning* was performed with both the constant-velocity and periodic-velocity modes for 10,000 training steps using the low-rank multivariate normal ("lrmn") guide.

Likewise, for the third sample of wild-type RPE1 cells (3,354 cells), *manifold-learning* was run on the "medium" gene set available in velocycle.utils, after filtering out genes with <=0.1 mean unspliced UMIs per cell or with <=0.3 mean spliced UMIs per cell (128 genes

remaining). *Manifold-learning* was performed using 3,000 training steps, and *velocity-learning* was performed with both the constant-velocity and periodic-velocity modes for 10,000 training steps using the low-rank multivariate normal ("lrmn") guide.

3.5.7. Experimental procedures

3.5.7.1. Cell culture

FUCCI-RPE-1 cells (see **Fig. 4**), a gift from Battich et al [62] were cultured at 37°C and 5% CO_2 in DMEM/F12 medium (Gibco 11320033) supplemented with 1% NEAA (Gibco 11140-035), 1% Penicillin/Streptomycin (Sigma Aldrich G6784) and 10% FBS (Gibco 10437-028).

Additional RPE-1 cells (see **Fig. 5**) were obtained from ATCC and cultured at 37°C, 20% O_2, and 5% CO_2 in DMEM/F12 medium (Gibco 21331-020) supplemented with 1% MEM NEAA (Sartorius 01-340-1B), 0.5% sodium pyruvate 1% penicillin/streptomycin/glutamine, and 10% FBS. Media was replaced daily and cells were passaged twice a week. RPE-1 cells were maintained in culture for at least two passages and confirmed to be free of mycoplasma.

3.5.7.2. scRNA-seq library preparations

For the preparation of scRNA-seq libraries, an experimental setup was designed to mimic the conditions used for live-cell imaging. FUCCI-RPE-1 cells (see **Fig. 3.4**) were seeded (7000 cells/cm²) in duplicate 2 days before collection. On collection day, cells were detached with trypsin, washed with PBS, counted and diluted to a cell concentration of 1000 cells/uL. Barcoded cDNA libraries were generated from single-cell suspensions using the 10x Genomics Chromium v3.1 dual-index system. The procedure was done in accordance with the manufacturer's instructions, with a goal of 4,000 cells per library. Samples were individually indexed and evenly pooled together. After quality control, libraries were sequenced on an Illumina HiSeq4000 platform, with a depth of approximately 300 million reads per sample, by the EPFL Gene Expression Core Facility (GECF).

Similarly, RPE-1 cells (see **Fig. 3.5**) were detached using Trypsin–EDTA solution A 0.25% (Biological Industries; 030501B) for 5 min at 37°C. Trypsin was neutralized with medium including 10% FBS and cells were centrifuged at 250 rcf for 5 min, followed by washing and resuspension in PBS with 0.04% BSA. The cell suspension was filtered with a 40-µm cell strainer to remove cell clumps. A cell viability percentage higher than 90% was determined by trypan blue staining. Cells were diluted to a final concentration of 700 cells per µl. scRNA-seq libraries were generated using the 10x Genomics Chromium v3.1 dual-index system. The procedure was done in accordance with the manufacturer's instructions, with a

goal of 3,000 cells per library. Samples were then indexed and sequenced on an Illumina NovaSeq 6000 platform by the EPFL Gene Expression Core Facility (GECF).

As with all publicly available datasets, raw fastq files were processed with CellRanger using the default human reference transcriptome to obtain count matrices. To obtain unspliced and spliced count matrices, we used velocyto version 0.17.17.

3.5.7.3. Live-image microscopy and cell tracking experiments with RPE1 cells

RPE-1 cells were seeded on glass bottom 6 well chamber slides (IBIDI) to reach 30% confluence after one day. Cells were then imaged on a PerkinElmer Operetta microscope under controlled temperature and CO_2 every 10.25 minutes using brightfield and digital phase contrast (DPC) with a 10x (0.35 NA) air objective, binning of 2 and the speckle scale set to 0 under non-saturated conditions. Cell division tracking was achieved by stacking time-course images and manually tracing cell movement and division with napari (Sofroniew et al., 2021). Between 20-25 RPE1 cells were tracked from 15 different fields of view by three different individuals (A.R.L., A.H., A.V.) for a total of 337 cells used to estimate a ground truth.

3.5.7.4. Cumulative EdU and p21 staining experiments

Cells were seeded on poly-L-lys-coated 24 well plates to reach 30% confluence after one day. After a day, 10 µM EdU (Invitrogen - #A10044) was added to the media, and cells were fixed at different time points after EdU addition: 30 min, 1h, 2h, 3h, 5h, 8h, 12h, 24h, 32h, 36h, 49h, 56h, 72h. For each time point, cells were fixed in 4% PFA for 10 minutes, washed twice with PBS, and processed for EdU (5-ethynyl-2'-deoxyuridine) detection according to the manufacturer's instructions (Click-iT EdU Alexa Fluor 647 Imaging kit from Invitrogen #C10340). Additionally, cells were permeabilized with 0.2% triton and stained overnight at 4°C with p21 Waf1/Cip1 (12D1) Rabbit mAb (Cell Signaling Technology #2947) and revealed with a secondary antibody conjugated to Alexa Fluor-488. After staining, cells were imaged on a Leica DMi8 (20x NA 0.8).

To quantify the signal intensities of p21 and EdU, we segmented nuclei in the DAPI channel using stardist (U. Schmidt et al., 2018). We obtained the average intensity for both signals per nucleus by sub-setting the corresponding channel using segmentation masks. The intensity of p21 was normalized per image (percentile-based, p_min=1, p_max=99.8), as its intensity profile was expected to be approximately constant in time; conversely, the intensity of EdU was not normalized as it was expected to increase with time. Thresholds were selected observing the (bimodal) signal distribution across nuclei in all timepoints.

First, to compute the time it takes, on average, for a cell to traverse through two consecutive S phases (tEdU), we applied the Nowakowski method (Di Bella et al., 2021; Nowakowski et al., 1989) on data collected at multiple time points for a total number of 678,204 cells. The Nowakowski method assumes a linear growth of EdU+ cells, until reaching a plateau where all cycling cells are positive for EdU. We obtained a linear fit of the growth and determined the x-value at the intersection between growth and plateau (A), the y-intercept of the linear fit (B), and the y-value of the plateau (GF). With these, we could compute tEdU as follows: tEdU = (B*A)/(GF-B) + A.

However, cells may on occasion exit the cycle to a G0 phase, and then re-enter at a later time (Krenning et al., 2022). To correct for this, we plotted the fraction of EdU+/p21+ cells among the p21+ population to estimate the G0 duration (tG0). We determined the tG0 to be equal to the time point at which fraction of EdU+/p21+ cells plateaued, after which no statistically significant changes were detected (Tukey's multiple comparison test). The corrected estimate for the cell cycle duration was finally calculated as: tc = tEdU - (%p21 * tG0), where %p21 corresponded to the mean fraction of p21+ cells.

3.5.8. Data Availability

The raw and processed scRNA-seq data in the RPE1 cell line that was newly generated for this study are available at GEO accession number **GSE250148.** All other scRNA-seq data used in this study were collected from previously published works (Aissa et al., 2021; Buettner et al., 2015; Capolupo, Khven, et al., 2022; La Manno et al., 2021; Replogle et al., 2022; Riba et al., 2022) and relied on the cell type annotations made by the original authors. Jupyter notebooks and other affiliated files to reproduce the results shown in this study are provided via a link available on our GitHub page: https://github.com/lamanno-epfl/velocycle/. Processed versions of all published data (including spliced-unspliced counts matrices) are also available at the above link. The simulated scRNA-seq datasets, processed scRNA-seq metadata for the new RPE1 samples, cell tracking data from live-image microscopy and cumulative EdU staining experiments are also available at the above link.

3.5.9. Code Availability

VeloCycle is implemented in Python and available as an open-source package on GitHub at https://github.com/lamanno-epfl/velocycle. *VeloCycle* can be installed from PyPi using the command *pip install velocycle* or via direct installation from the GitHub page using the command *pip install git+https://github.com/lamanno-epfl/velocycle*.git@main. Python version 3.8 or newer is required. Source code, installation instructions, tutorials, and a file

containing all required package dependencies are also available on GitHub. Additional code and notebooks to reproduce the results of this study are available via the link provided on the GitHub page.

3.5.10. Acknowledgements

This project has been made possible in part thanks to Chan Zuckerberg Initiative grants number 2022-249212 and 2019-002427. G.M.L. received support from the Swiss National Science Foundation (SNSF) grant PZ00P3_193445. F.N. received support from the SNSF grant #310030B_201267 and the EPFL. L.P. is partially supported by the National Human Genome Research Institute (NHGRI) Genomic Innovator Award (R35HG010717). We thank members of the La Manno, Naef, Pinello, Williams and Castelo-Branco labs for their generous feedback and discussions on the project, particularly Daniil Bobrovskiy, Cameron Smith, Qian Qin, Eli Bingham, Luise Seeker, Nadine Bestard, Mukund Kabbe, and Fabio Baldivia Pohl. We also thank the entire EPFL Gene Expression Core Facility (GECF) for their assistance with scRNA-seq experiments.

3.5.11. Author Contributions

A.R.L. developed the idea, designed, implemented, and refined the model framework, analyzed the scRNA-seq and time-lapse microscopy data, created the figures, and wrote the manuscript. M.L., L.T., and C.D. participated in development of the idea and model formulation. A.H., A.V., I.K., and H.C. performed scRNA-seq experiments. A.H. and A.V. also conducted time-lapse microscopy and cumulative EdU experiments, with image processing analysis aid from A.D.M. P.M.A. helped test phase estimation approaches. L.P. helped refine the idea and related analyses. F.N. and G.L.M. developed the idea, supervised the project, and wrote the manuscript. All co-authors read and approved the manuscript.

3.6. Supplementary Materials

3.6.1. Supplementary Figures

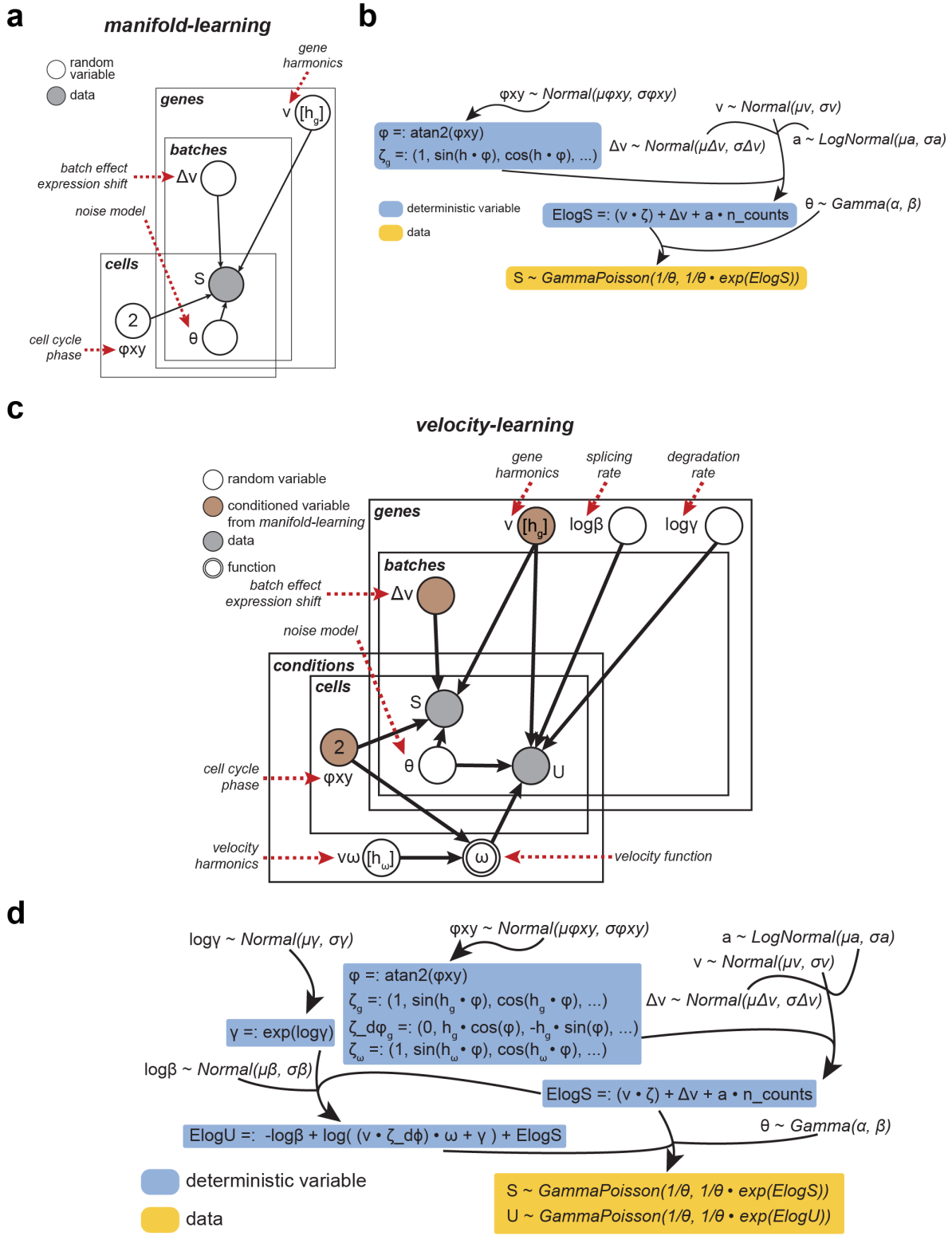


Figure S3.1. Plate diagram and mathematical formulation of *VeloCycle* framework for *manifold-learning* and *velocity-learning*. (a) Plate notation diagram of the *manifold-learning* procedure used to infer manifold coordinates (ϕ) and geometry (v) given raw spliced count data. No k-nearest neighbor smoothing is performed. The model assigns each cell to a phase along the cell cycle (ϕ) and fits a set

of harmonics (v) for each individual gene using a Fourier series. (b) Formulaic representation of the *manifold-learning* procedure shown in (a). Spliced counts (S) are defined as the expectation (ElogS) plus noise, modeled after a negative binomial distribution. (c) Plate notation diagram of the complete *velocity-learning* procedure. Nodes indicate a model variable (white: random variable; gray: observed data; brown: conditioned variable from *manifold-learning*) and arrows indicate dependency. Each plate (genes, cells, conditions, and batches) signal independence and contain variables that are of the same dimensions. (d) Mathematical representation of the *velocity-learning* procedure shown in (c). Blue-boxed variables are deterministic and computed from latent variables; yellow-boxed variables are conditioned on observed data.

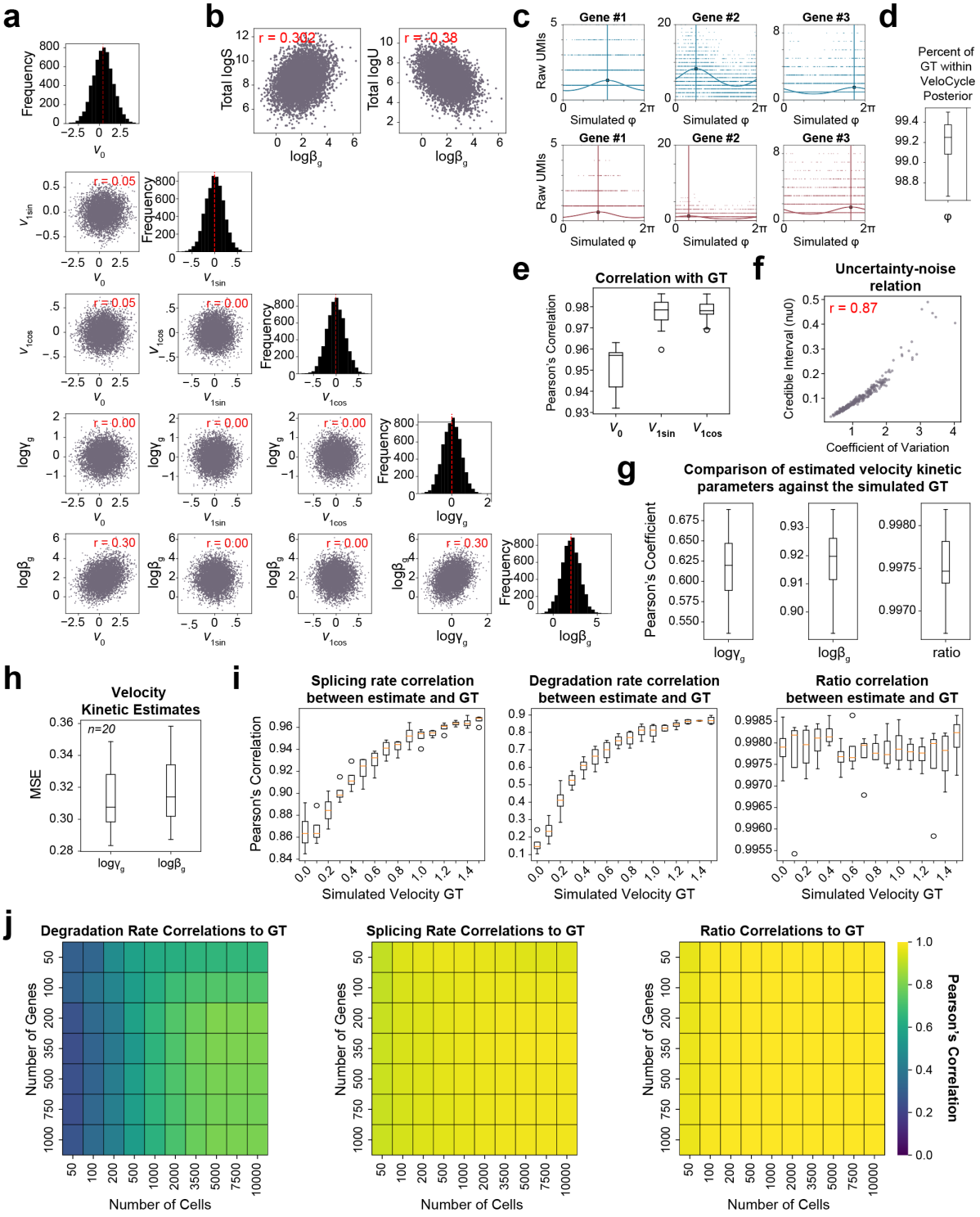


Figure S3.2. Data generated with structured simulations assists in validation of *VeloCycle*. (a) Scatter plot grid of the structured correlation between gene harmonics coefficients (v_0, v_{1sin}, v_{1cos}) and kinetic parameters ($\log\beta_g$, $\log\gamma_g$) in simulated ground truth (GT) data. Histograms (on diagonal) of the frequency distribution for each simulated latent variable. **(b)** Scatter plots of the simulated data correlation structure between splicing rate ($\log\beta_g$), total spliced (logS), and unspliced (logU) UMI counts. **(c)** Scatter plots of example simulated gene fits for spliced (top; blue) and unspliced (bottom; red) UMIs. Solid curved lines represent gene fits, and vertical lines indicate the phase of peak expression for each gene. **(d)** Box plot of the percent of GT phases within the posterior uncertainty interval estimated, across 20 independently-simulated datasets each containing 3,000 cells and 300 genes. **(e)** Box plots of the

mean circular correlation coefficient, across 300 genes, for estimated gene harmonic coefficients obtained compared to the GT. **(f)** Scatter plot of the gene-wise coefficient of variation (a measure of noise) and the credible interval obtained for the gene harmonic coefficient v_0. **(g)** Box plots of the mean Pearson's correlation coefficient for all genes, across 20 independently-simulated datasets, between the estimated and simulated GT for the degradation rate ($\log\gamma_g$), splicing rate ($\log\beta_g$), and velocity kinetic ratio ($\log\gamma_g$-$\log\beta_g$). **(h)** Box plots of the mean squared error (MSE) for $\log\gamma_g$ and $\log\beta_g$ against the simulated GT for 20 datasets in (g). **(i)** Box plots of the mean Pearson's correlation coefficient between estimated and GT values for $\log\beta_g$, $\log\gamma_g$, and velocity kinetic ratio, for all genes across four independently-simulated datasets with 16 different velocity GT between 0.0 and 1.5. **(j)** Heatmaps showing the correlation between estimated and GT values for $\log\beta_g$, $\log\gamma_g$, and velocity kinetic ratio using varying numbers of cells and genes. Pearson's correlation coefficients (red) are indicated in each scatter plot of (a), (b), and (f). Each purple dot represents a single gene in (a), (b), and (f).

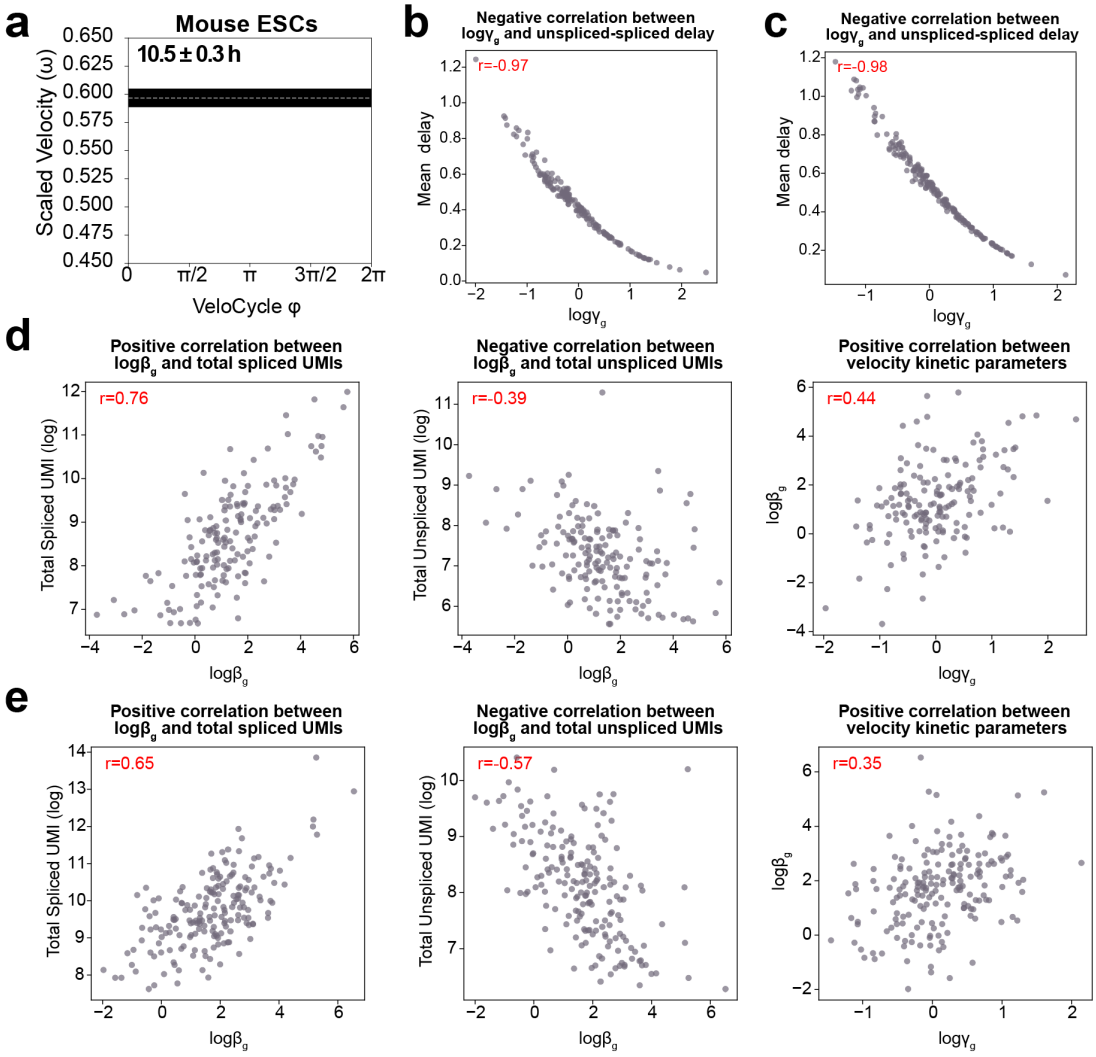


Figure S3.3. ***VeloCycle*** **accurately measures phase and speed of the cell cycle across species.** **(a)** Posterior estimate plot of cell cycle velocity inferred with *velocity-learning* on mouse ESCs (Riba et al., 2022). White dashed lines represent the mean of 500 posterior samples; black bars indicate the full posterior interval. **(b)** Scatter plot of the relationship between degradation rate ($\log\gamma_g$) and the average unspliced-spliced delay in human fibroblasts. **(c)** Scatter plot of the relationship between degradation rate ($\log\gamma_g$) and the average unspliced-spliced delay in mouse ESCs. **(d)** Scatter plots of the relationships among splicing rate ($\log\beta_g$), degradation rate ($\log\gamma_g$), and total UMI counts (spliced and unspliced) in human fibroblasts. **(e)** Scatter plots of velocity kinetic parameter relationships for mouse ESCs, as in (d). Pearson's correlation coefficients (red) are indicated in the top right of plots in (b-e).

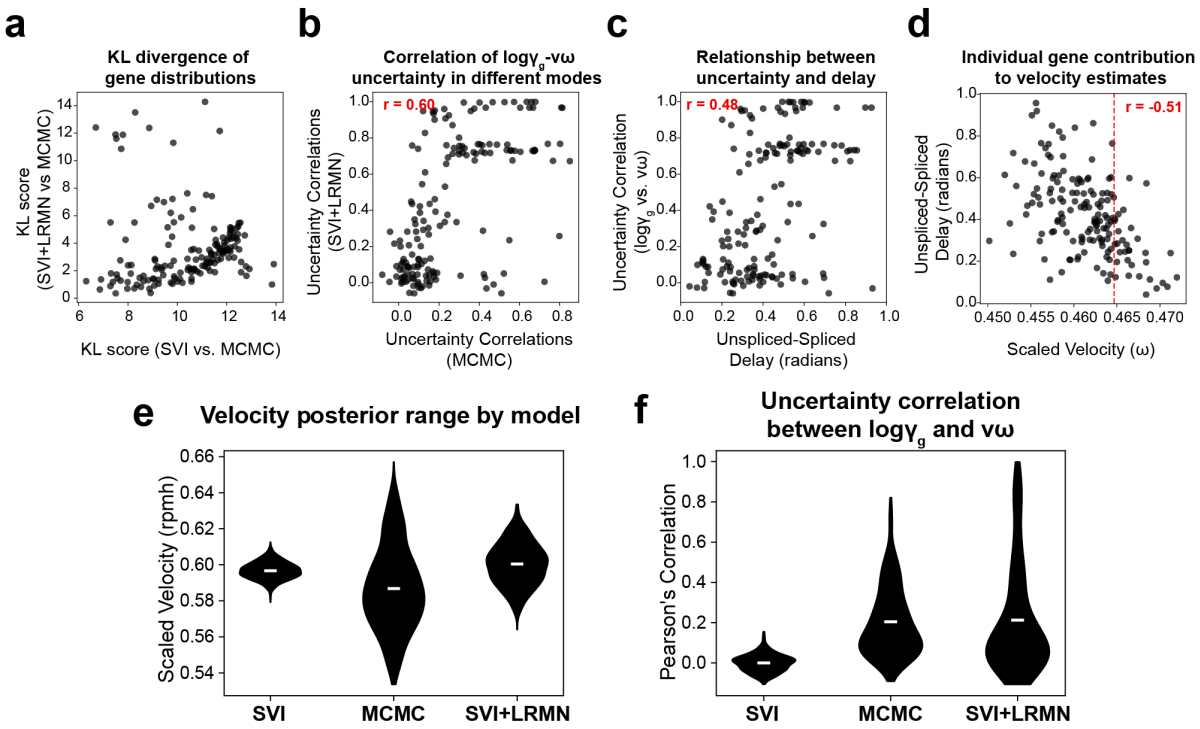


Figure S3.4. A structured variational distribution yields better velocity uncertainty estimates and reveals relationships among gene kinetic parameters. (a) Scatter plot of gene-wise Kullback–Leibler (KL) divergence comparing uncertainty distributions between SVI and MCMC (x-axis) and SVI+LRMN and MCMC (y-axis). A lower KL divergence indicates a greater overlap between the two distributions. **(b)** Scatter plot between the gene-wise $\log\gamma_g$-$v\omega$ uncertainties computed from the posterior of MCMC or SVI+LRMN. **(c)** Scatter plot between unspliced-spliced peak expression delay (radians) and the $\log\gamma_g$-$v\omega$ uncertainty correlation, both obtained using the SVI+LRMN velocity model. **(d)** Scatter plot between the scaled velocity and the unspliced-spliced delay for a leave-one-out estimation approach with *VeloCycle*. Each dot is positioned on the x-axis at the velocity estimate obtained when removing a particular gene (n=160) from the gene set. Each dot is positioned on the y-axis at the position of the unspliced-spiced delay (in radians) for that removed gene. **(e)** Violin plots of the scaled velocity for mouse embryonic stem cells (mESCs), comparable to Fig. 3.5c. **(f)** Violin plots of the Pearson's correlations between the degradation rate ($\log\gamma_g$) and angular speed ($v\omega$) posteriors across all 189 genes for mESCs, comparable to Fig. 5d. Pearson's correlation coefficients are indicated in red in (b), (c), (d).

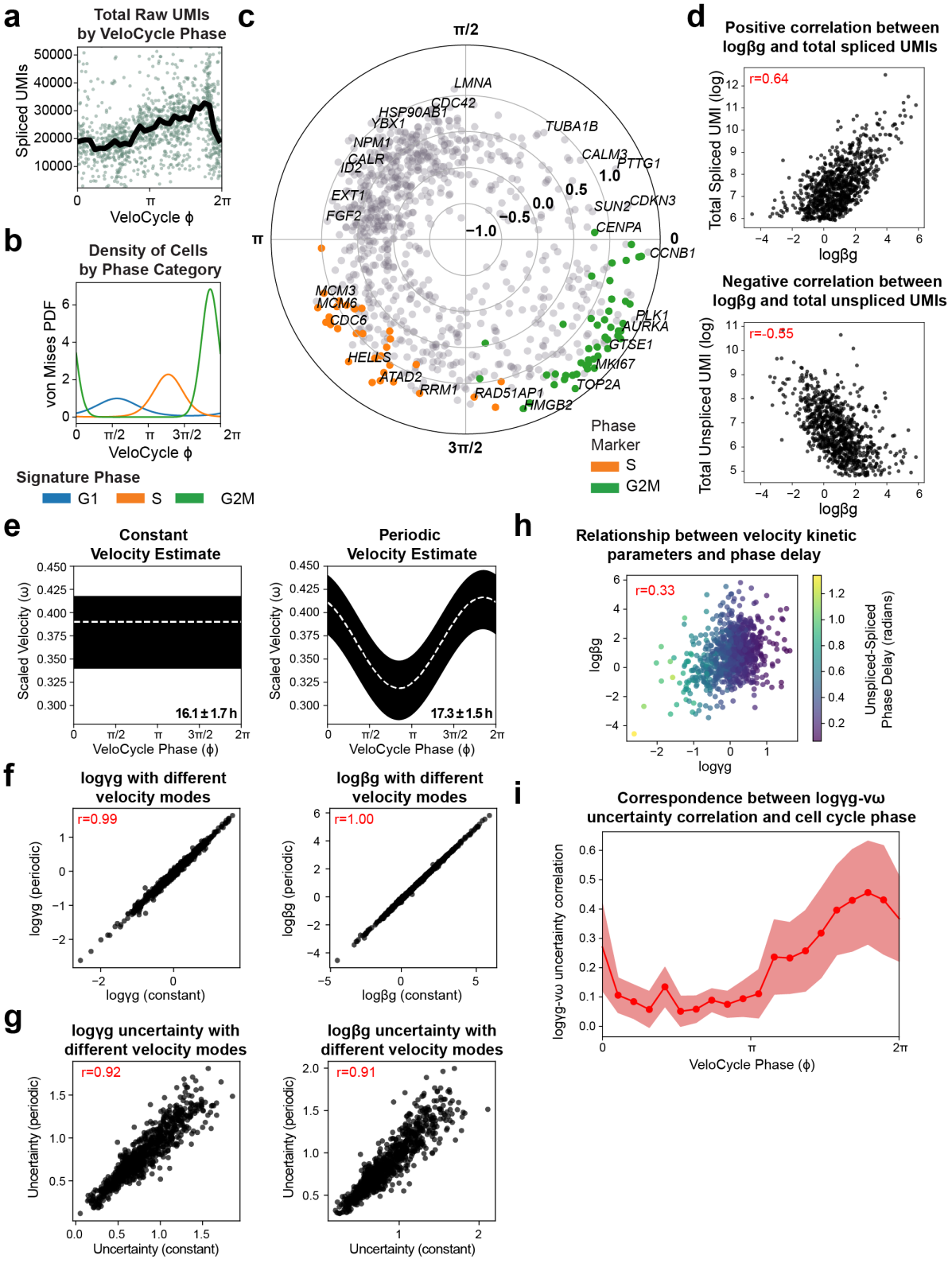


Figure S3.5. *VeloCycle* coupled with live-cell imaging enables experimental validation of cell cycle speed. (a) Scatter plot of total raw spliced UMI counts by continuous cell cycle phase estimated with *VeloCycle* for 1,222 dermal human fibroblasts (dHFs). Black line indicates the binned mean. **(b)** Probability density plot along the *VeloCycle* phase estimation for cells in (a) stratified by categorical phase assignment (G1, S, and G2M) obtained with scanpy. **(c)** Phase space polar plot indicating the phase of peak expression and amplitude for 876 cycling genes used to learn the manifold for dHFs in (a). Each dot represents a gene; genes colored orange (S) or green (G2M) represent marker genes

used in traditional categorical cell cycle phase assignment. **(d)** Scatter plots of the relationship between splicing rate and total spliced counts (top) and splicing rate and total unspliced counts (bottom) on a gene-wise basis. **(e)** Posterior estimate plot of constant (left) and periodic (right) velocity estimates obtained for data in (a) using a medium-sized gene set (Riba et al., 2022). **(f)** Scatter plots of degradation rates (left) and splicing rates (right) obtained using either the constant (x-axis) or periodic (y-axis) models of velocity estimation. **(g)** Scatter plots of degradation rate (left) and splicing rate (right) posterior uncertainty obtained from 500 posterior samples using either the constant (x-axis) or periodic (y-axis) models. **(h)** Scatter plot of the degradation and splicing rates obtained with the SVI+LRMN model on data in (a-f). Gene-wise dots are colored by the unspliced-spliced phase delay. **(i)** Binned plot of the Pearson's correlation coefficients between gene-wise degradation rate and velocity posterior uncertainties on dHFs using the SVI+LRMN model of VeloCycle. Pearson's correlations coefficients are indicated in red text in (d-g).

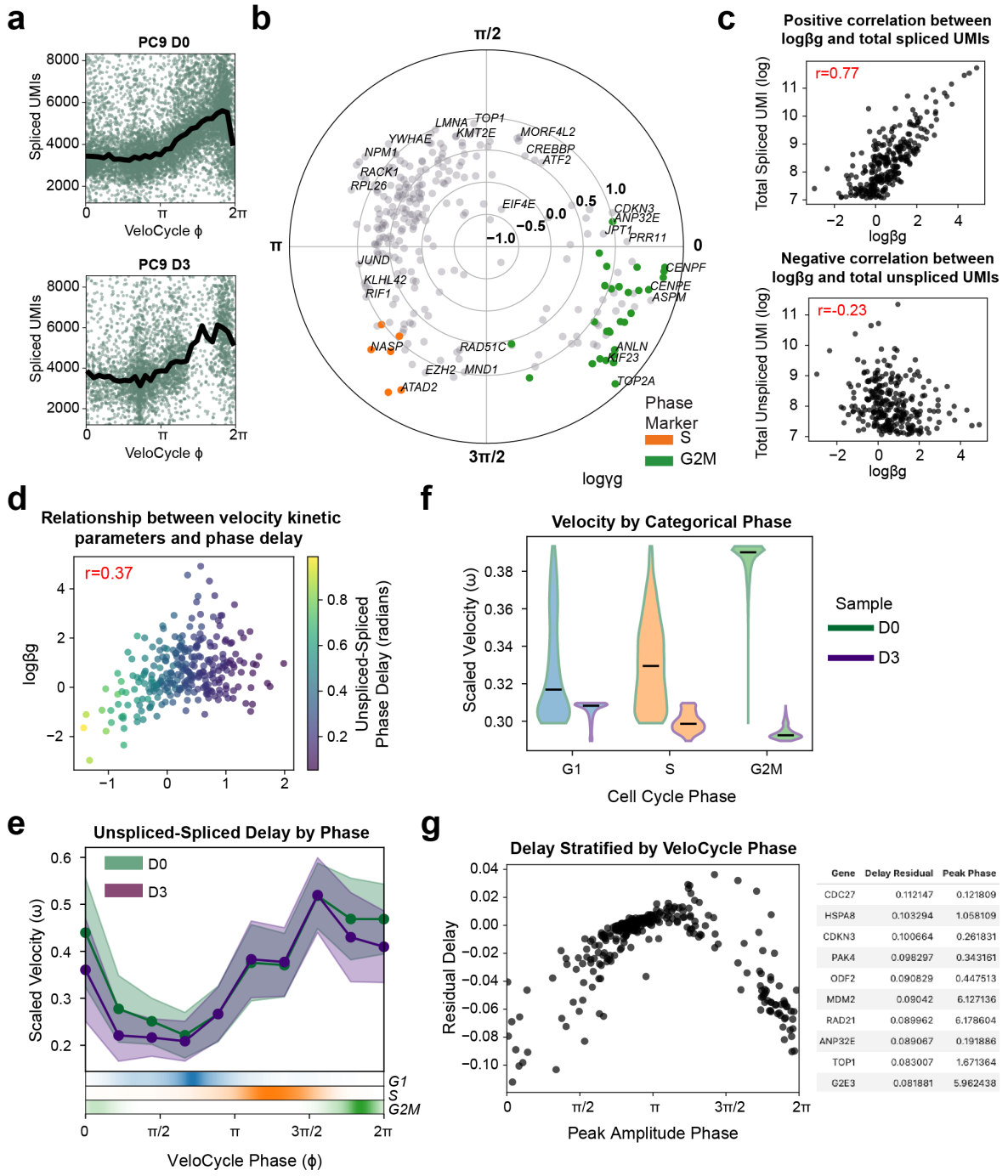


Figure S3.6. Statistical credibility testing of RNA velocity estimates enables characterization of the effect of erlotinib treatment on lung adenocarcinoma cell line treatment. (a) Scatter plot of total raw spliced UMI counts by continuous cell cycle phase estimated with *VeloCycle* for PC9 lung adenocarcinoma cell line populations before (top; D0; 9,927 cells) and after (bottom; D3; 3,943 cells). Black line indicates the binned mean. **(b)** Phase space polar plot indicating the phase of peak expression and amplitude for cycling genes used to learn the manifold for cells in (a). Each dot represents a gene; genes colored orange (S) or green (G2M) represent marker genes used in traditional categorical cell cycle phase assignment. **(c)** Scatter plots of the relationship between splicing rate and total spliced counts (top) and splicing rate and total unspliced counts (bottom) on a gene-wise basis. **(d)** Scatter plot of the degradation and splicing rates obtained with the SVI+LRMN model. Gene-wise dots are colored by the unspliced-spliced phase delay. **(e)** Gene-binned delay between maximum unspliced-spliced expression (in radians) for D0 and D3 samples. **(f)** Violin plots of scaled velocity

estimates for D0 and D3, stratified by categorical cell cycle phase. Black horizontal lines indicate the mean. **(g)** Left: scatter plot of peak gene amplitude and residual unspliced-spliced delay (D3-D0) for 273 genes. Right: top 10 differentially delayed genes in D0 versus D3.

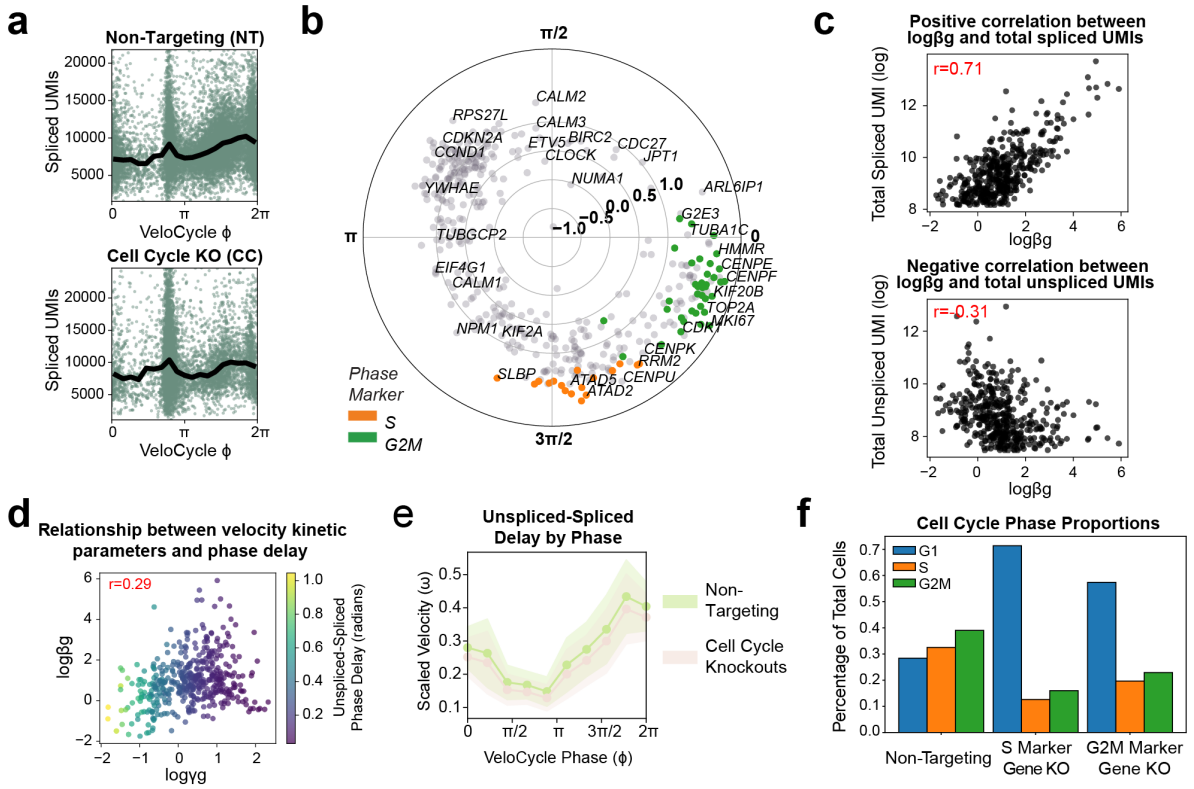


Figure S3.7. Cell cycle velocity estimation on non-targeting and grouped cell cycle knockout stratifications of RPE1 cells following genome-wide Perturb-seq. (a) Scatter plots of total UMIs along the manifold-learning cell cycle phase for non-targeting (NT; top) and cell cycle knockout (CC) strata of genome-wide Perturb-seq data from Fig. 8. **(b)** Phase space polar plot indicating the phase of peak expression and amplitude for 426 cycling genes used to learn the manifold for cells in (a). **(c)** Scatter plots of the relationship between splicing rate and total spliced counts (top) and splicing rate and total unspliced counts (bottom) on a gene-wise basis. **(d)** Scatter plot of the degradation and splicing rates; gene-wise dots are colored by the mean unspliced-spliced phase delay. **(e)** Gene-binned delay between maximum unspliced-spliced expression (in radians) for NT and CC samples. **(f)** Bar plots of categorical cell cycle phase proportions as a percentage of total number of cells, stratified by Perturb-seq non-targeting, S phase marker gene, and G2M marker gene conditions.

3.6.2. Supplementary Tables

Variable	Description	General Name	Training Step	Dimensions
φ_{xy}	cell cycle phase	Manifold coordinates	*manifold-learning*	(cell)
v	Fourier coefficients for the genes	Manifold geometry	*manifold-learning*	(gene, harmonics)
Δv	batch-specific expression offset	Data-specific noise in manifold geometry	*manifold-learning*	(batch, gene)
$shape\_inv$	spliced negative binomial noise	Measurement noise	*manifold-learning*	(gene)
$\log\ \beta_g$	log splicing rate	Velocity kinetics	*velocity-learning*	(gene)
$\log\ \gamma_g$	log degradation rate	Velocity kinetics	*velocity-learning*	(gene)
$v\omega$	Fourier coefficients for the angular speed	Velocity function	*velocity-learning*	(condition, harmonics)

Table 3.1. Overview of *VeloCycle* latent variables.

Please refer to the preprint for supplemental tables **Table 3.2** and **Table 3.3**:

https://doi.org/10.1101/2024.01.18.576093

4. Charting recurring cell lipid-state transitions with lineage leaf-state Markov analysis reveals the stability of metabolic configurations

Alex R. Lederer[1], Laura Capolupo[2], Gioele La Manno[1], Giovanni D'Angelo[2]

[1] Brain Mind Institute, Faculty of Life Sciences, EPFL, Lausanne Switzerland
[2] Interfaculty Institute of Bioengineering and Global Health Institute, EPFL, Lausanne Switzerland

This work was published as part of a research article entitled "Sphingolipids control dermal fibroblast heterogeneity" in *Science* (Capolupo, Khven, et al., 2022) in April 2022. My contribution as the third author to the larger publication is the focus of this thesis chapter, and the complete article is available at: https://doi.org/10.1126/science.abh1623.

4.0. Preface

In this chapter, I describe a project carried out in collaboration with the lab of Prof. Giovanni D'Angelo at the EPFL. Some of the text in following sections is adapted from Supplementary Note 1 and the results section of the published paper. The figures are also adapted from the publication.

The experimental work presented here, including cell culture, time-lapse microscopy, MALDI-MSI, and toxin staining was performed by Dr. Laura Capolupo (a former PhD student in the D'Angelo lab). Laura also performed image segmentation and cell tracking. I conceptualized and implemented the CELLMA model, assisted with data processing, and helped with writing the results during manuscript revisions. Other author contributions are not directly related to CELLMA and can be found under the "Author Contributions" section of the article above. The author list of the original publication is as follows: Laura Capolupo, Irina Khven, Alex R. Lederer, Luigi Mazzeo, Galina Glousker, Sylvia Ho, Francesco Russo, Jonathan Paz Montoya, Dhaka R. Bhandari, Andrew P. Bowman, Shane R. Ellis, Romain Gulet, Olivier Burri, Johanna Detzner, Johannes Muthing, Krisztian Homicsko, François Kuonen, Michel Gilliet, Bernhard Spengler, Ron M. A. Heeren, G. Paolo Dotto, Gioele La Manno, Giovanni D'Angelo.

4.1. Synopsis

Charting the differentiation trajectories of progenitor cells into mature lineages is of major interest to the systems biology community. Consequently, most models of single cell state transition, despite being ergodic in principle, imply or expect a directionality in which each state is visited at most once. These methods are not applicable to studying recurrent dynamics, where a cell visits a state multiple times and in an acyclical manner. These situations require time resolved information, yet most of the field has focused on gene expression snapshot data. Here, we propose a new computational approach to tackle these limitations by coupling live imaging and omics recordings. Using time-lapse microscopy, toxin stainings, and endpoint matrix-assisted laser desorption/ionization mass spectrometry imaging (MALDI-MSI), we measure lipid content at a single-cell spatial resolution and model confirmation switches with cell-state transition estimation by lineage leaf-state Markov analysis (CELLMA). We apply CELLMA to dermal human fibroblasts, which transit among various sphingolipid configurations to play diverse roles in wound healing and proliferation. We observe that dominant, stable lipid-states are propagated across cell generations and correspond to phenotypic states called lipotypes. CELLMA holds future potential to estimate single-cell state transition probabilities for hundreds of lipids measured with MALDI-MSI.

4.2. Introduction

Computational methods that capture the progression of cells along a one-directional trajectory with single-cell transcriptomic measurements are regularly applied to study differentiation and embryonic development. Furthermore, models have been devised to describe scenarios in which cells revisit past states, even multiple times, in a circular manner (Chervov & Zinovyev, 2022). Two well-known periodic biological phenomena are the cell cycle and circadian rhythms. The cell cycle regulates the replication of genetic material and cell division (Matson & Cook, 2017), whereas the 24-hour circadian clock is responsible for modulating numerous physiological changes in the mammalian body (Talamanca & Naef, 2020). Recent work has sought to estimate cell state transitions in these periodic processes and characterize their behaviors (**Chapter 3**) (Talamanca et al., 2023).

However, there is yet another class of cell transitions that feature recurring state dynamics, but in a non-cyclical manner. In these systems, cells may transition among multiple cell states, each of which is responsible for performing a specific functional role in tissue. These different states possess varying degrees of stability and, as a result, a cell may not always pass through all possible states. Thus, the observed cell state transitions may repeat, but not always in the same order. Unfortunately, existing computational methods are poorly

suited to address these types of recurring cell dynamics, which do not occur in embryonic development but rather in biological systems at homeostasis.

One example of such a transition process is the maturation of blood cells in the bone marrow. There, hematopoietic stem cells respond to both internal and external cues, oscillating among states of self-renewal, differentiation, and maturation in a highly coordinated manner (Liggett & Sankaran, 2020; Ye et al., 2017). This approach ensures both production of mature blood cell types, the rate of which may change depending on the immune state of the individual, as well as maintenance of a source progenitor pool (Haghverdi & Ludwig, 2023)

Another example is dermal human fibroblasts (dHF), which fluctuate among fibrogenic, fibrolytic, and immune states to facilitate wound healing, remodel the extracellular matrix, and promote tissue proliferation. Changes in the proportion of dHF subpopulations are thought to be the combined result of cell-autonomous factors, external signals, and cell lineage (Adler et al., 2020; Driskell & Watt, 2015; Philippeos et al., 2018; Rognoni et al., 2018). However, the rules governing this established heterogeneity are not fully understood.

Interestingly, lipids have been demonstrated to modulate the differentiation of stem cells in the skin (Cliff & Dalton, 2017; D. Russo, Capolupo, et al., 2018). Whether metabolic switches involving lipids can define cell state transitions and facilitate fibroblast plasticity has not been well-studied. This is in part due to technical limitations of evaluating lipid composition at a single-cell resolution, which has only been pursued by a few studies (Denz et al., 2017; Frechin et al., 2015; D. Russo, Della Ragione, et al., 2018; Snijder et al., 2009).

Fortunately, mass spectrometry (MS) techniques now have enough sensitivity to enable single-cell lipidomics (Rappez et al., 2021; Thiele et al., 2019; Zenobi, 2013). In particular, matrix-assisted laser desorption/ionization mass spectrometry imaging MALDI-MSI provides coverage of the lipid mass-to-charge range, causes minimal fragmentation, and has reached a spatial resolution compatible with single-cell analysis while maintaining mass resolution and accuracy (Bhandari et al., 2020; Dueñas et al., 2017; Kompauer et al., 2017; Niehaus et al., 2019; Norris & Caprioli, 2013; Schober et al., 2012; Sugiyama et al., 2018; Zavalin et al., 2012). Therefore, these approaches can be applied to characterize lipid transitions and, ultimately, to develop a framework for inferring transition probabilities.

Here, we assess whether recurring dHF cell states can be described using lipids and then estimate cell transition probabilities among those lipid-defined states. To accomplish this, we propose the use of time-lapse microscopy and endpoint lipid-state assignment to represent cell transitions by lineage leaf-based Markov analysis. We apply our model to dHFs and identify the stability of various sphingolipid configurations, showing that lipid-states are propagated across cell generations.

4.3. Results

4.3.1. Toxin stainings describe lipid configurations in dermal human fibroblasts

In order to design a computational model to infer cell state transition probabilities, we first sought a reliable means to discriminate between lipid configurations. We previously evaluated the single cell lipidomes of 257 dHFs measured by MALDI-MSI, showing that lipid distributions varied among cells in culture (Capolupo, Khven, et al., 2022). However, we sought to further validate this finding using an orthogonal method that we could then apply to unambiguously categorize cells into lipid state classes.

Therefore, we employed fluorescently labeled bacterial toxins to stain cells according to different sphingolipid head groups: Shiga toxin 1a (ShTxB1a) binds to trihexosylceramides (Gb3) (Jacewicz et al., 1986), Shiga toxin 2e (ShTxB2e) binds to Gb3 and globosides (Gb4) (Müthing et al., 2009), and Cholera toxin B (ChTxB) binds the ganglioside GM1 (Heyningen S Van, 1974). The staining patterns of these toxins (**Fig. 4.1A**) reflected the variability observed by MALDI-MSI (**Fig. 4.1B**). ShTxB1a staining correlated best with Gb3 levels, and ShTxB2e staining correlated well with Gb3 and Gb4 levels, whereas neither correlated with sphingomyelin (SM) levels. ChTxB staining is a proxy for the levels of GM1 (Heyningen S Van, 1974), a sphingolipid not detected by MALDI-MSI, and did not correlate with any of the sphingolipids detected by mass imaging (**Fig. 4.1C-D**). These findings highlight the capacity of bacterial toxins to capture sphingolipid heterogeneity in dHFs.

We next classified dHFs based on their toxin stainings into ChTxB$^+$, ShTxB1a$^+$, ShTxB2e$^+$, ShTxB1a$^+$/2e$^+$, triple$^+$, and “other” (i.e., all other configurations) (**Fig. 4.1E**). When examining features associated with these categories, we observed that ShTxB1a$^+$/2e$^+$ and triple$^+$ cells were larger than ChTxB$^+$ and ShTxB2e$^+$ cells and that ShTxB1a$^+$/2e$^+$ had a more complex shape than ChTxB$^+$ cells (**Fig. 4.1F**). We also considered the cell-to-cell variability associated with exo/endocytic organelles (Liberali et al., 2014), where sphingolipid production and turnover take place (Hannun & Obeid, 2018). We determined that ShTxB1a$^+$/2e$^+$ dHFs have an expanded early endosomal compartment compared to other configurations, with ChTxB$^+$ dHFs showing an opposite phenotype. Similar, although less striking, changes were observed when looking at coat protein complex I (COPI) vesicles and at the Golgi complex (**Fig. 4.1F-G**). These findings led us to conclude that distinct lipid metabolic configurations, which we term “lipotypes”, exist in dHFs. These lipotypes can be identified using ChTxB, ShTxB1a, and ShTxB2e toxin stainings, correspond to specific phenotypes of cell shape and size, and possess unique endocytic and secretory statuses.

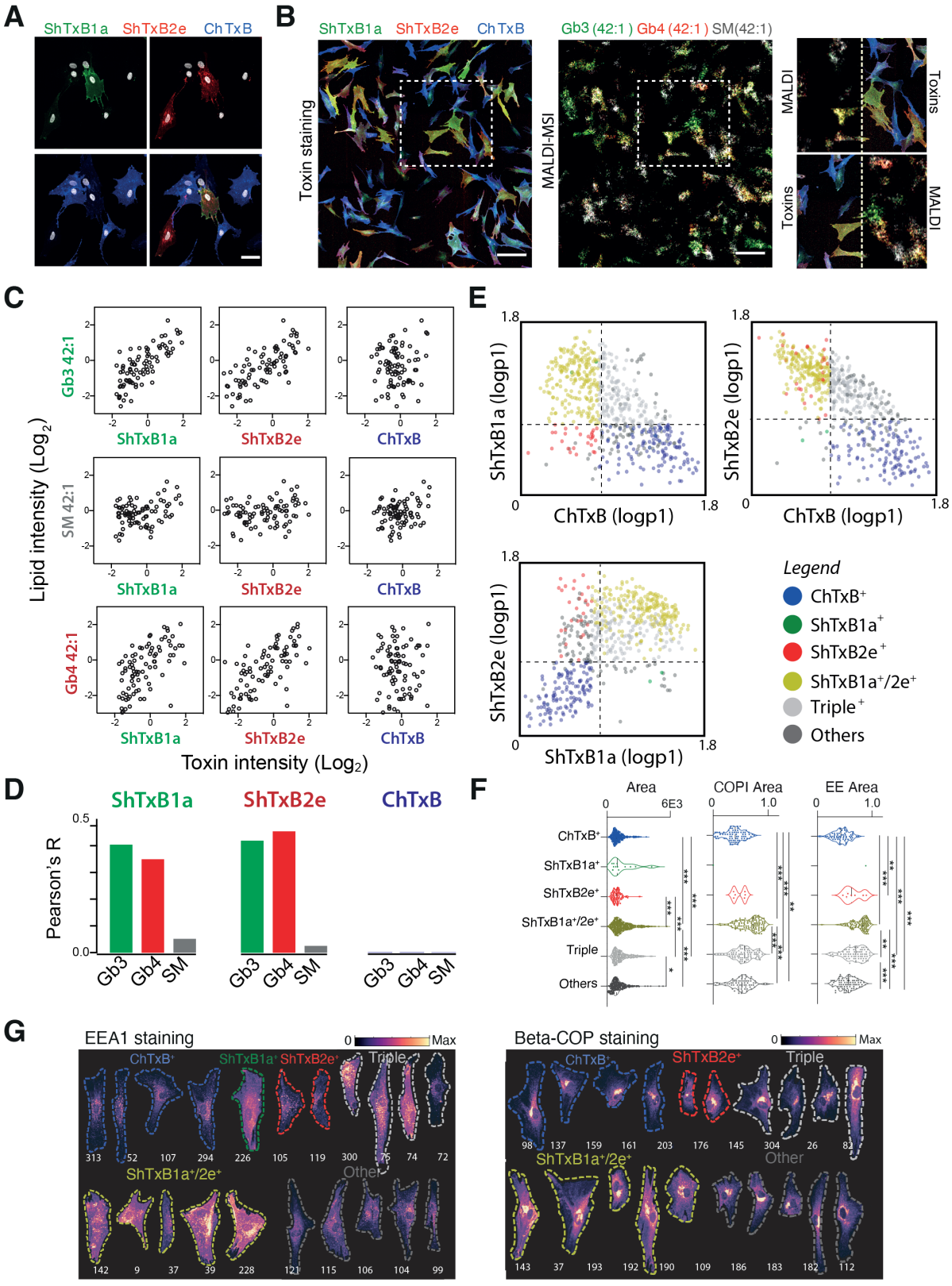


Figure 4.1. Identification of dHF lipotypes by MALDI-MSI and toxin staining. Adapted from Fig. S2, S3, and 3 of Capolupo et al 2022. **(A)** Confocal micrographs showing cells stained with bacterial toxins ShTxB1a (green), ShTxB2e (red), ChTxB (blue), and Hoechst (gray) for nuclei. Scale bar, 50µm. **(B)** Side-by-side comparison of toxin staining and MALDI-MSI acquisition on the same cells. First, cells were stained with bacterial toxins as in (A), and images were acquired by confocal microscopy (left panel). Then, MALDI-MSI (2m µm^2 / pixel) was performed in the same cells (center panel). Mass images

(320 x 320 pixels of complex sphingolipids [SM (42:1), Gb3 (42:1), and Gb4 (42:1)] are shown. Scale bar, 200 µm. **(C)** Scatter plots comparing toxin staining intensity and lipid abundance as determined by MALDI-MSI. Data were log2 transformed and centered. **(D)** Bar plot summarizing the Pearson's coefficient (R) for each toxin-lipid couple shown in (C); n=88. **(E)** Scatter plots of normalized fluorescence intensity for each toxin. Cells were stained using bacterial toxins as in (A) and categorized according to their lipid composition. Populations are colored as shown in the legend. **(F)** Cells were stained with bacterial toxins as in (A) and with antibodies against Beta-COP (COPI vesicles) or EEA1 [early endosomes (EEs)], and images were acquired by confocal microscopy. Normalized fluorescence intensities of toxin and organelle marker stainings of single cells were used to analyze the correlation between lipotypes and cell area and with exo/endocytic organelles. Data are shown as violin plots. *P < 0.05, **P < 0.01, ***P < 0.001, ordinary one-way ANOVA. **(G)** Representative cells stained for EEA1 or Beta-COP and classified according to their lipotypes.

4.3.2. Paired cell lineage reconstruction and endpoint toxin staining enable modeling of lipotype transitions

Recognizing the existence of lipotypes in dHFs under steady state conditions, we considered the relationships between those states. Given the technical limitations of single-cell omics approaches to temporally resolve lipid-defined state transitions, we established an experimental technique of time-lapse microscopy, which tracks cell movement and divisions, combined with endpoint toxin stainings (**Fig. 4.2A**). We hypothesized that by reconstructing the lineage relationships between cells, we could deduce whether particular lipotypes were transmitted across generations.

We monitored dHFs by time-lapse microscopy for 42 hours, with an image acquired every 20 minutes, followed by toxin staining after fixation (**Fig. 4.2A; Methods 4.5.7-4.5.8**). After cell segmentation and lineage reconstruction using Cellpose and TrackMate (**Methods 4.5.9**), we obtained a total 1,516 leaf cells with lipotype assignment grouped into 591 clades (**Fig. 4.2B-C, S4.1A-C**). To model the relationships between cells, we assumed that daughter cells inherit the same lipotype state from their mother at the moment of division and that cells with no apparent familial connection (i.e., from different clades) are independent (**Methods 4.5.3**).

Analysis of individual toxin levels indicated a strong correlation between pairs of sister cells (**Fig. 4.2D**). Moreover, lineage-related cells had a higher probability of sharing the same lipotype than expected by chance (**Fig 4.2E and S4.1D**), suggesting that lipid configurations can be propagated across cell generations and that lipotypes are long-term memory states.

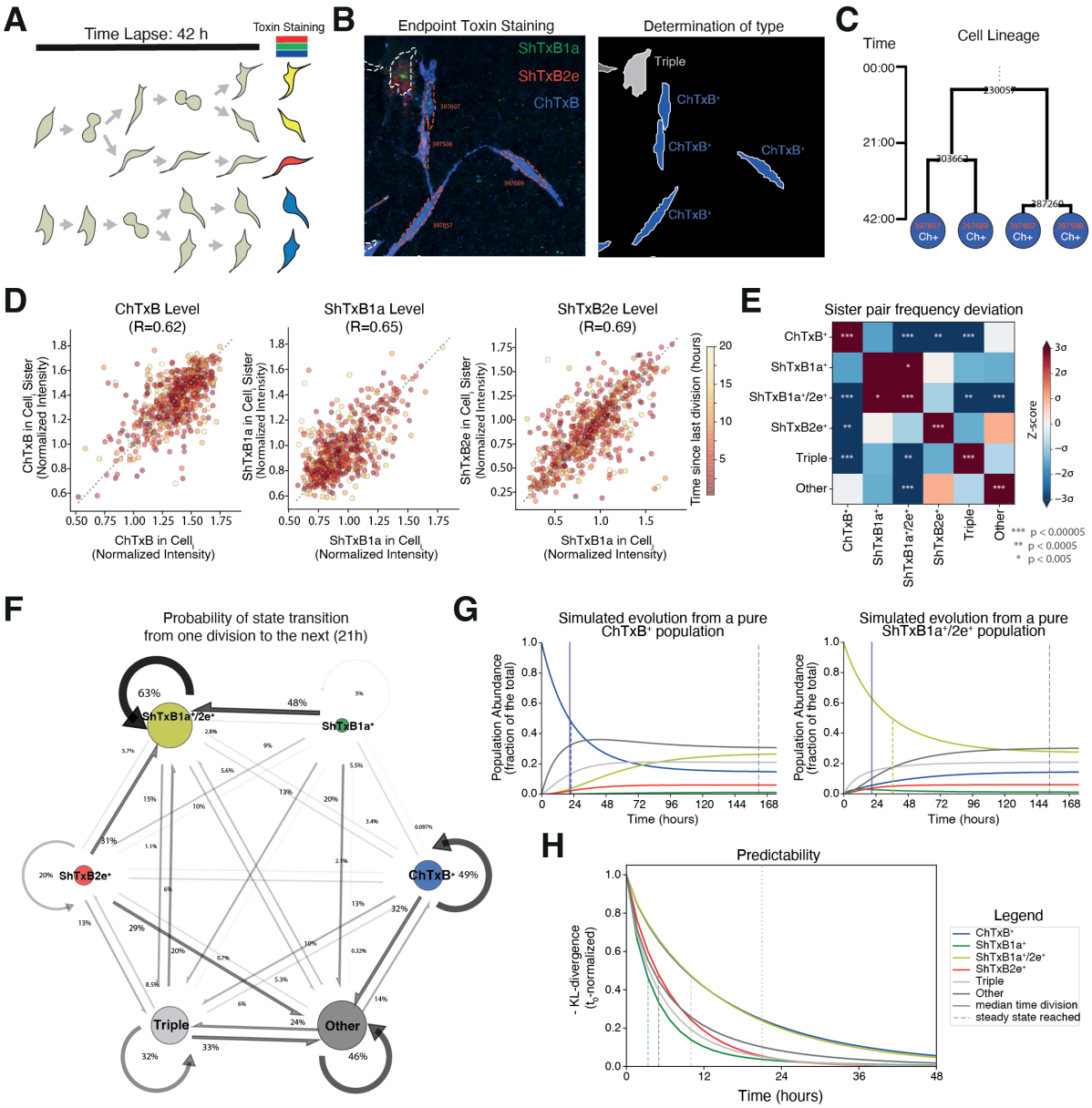


Figure 4.2. Lipotype transition estimations by lineage leaf-state Markov analysis. Adapted from Fig. 3 of Capolupo et al 2022. **(A)** Schematic representation of dHF cell lineage tracking. **(B)** Representative confocal micrograph of toxin-stained dHFs before (left) and after (right) segmentation with Cellpose. Segmented cell colors correspond to the different lipotypes. **(C)** Lineage reconstruction for the cells illustrated in (B) as inferred using TrackMate (**Methods 4.3.9**). **(D)** Correlation plots of normalized ChTxB, ShTxB1a, and ShTxB2e intensities between daughter cells at the time course end point. Dots are colored by the number of hours after mother cell division. **(E)** Heatmap of frequencies for two lipotypes occurring in two sister cells colored by z-score. Positive deviation from zero indicates an increased observed frequency of the sister-lipotype combination compared with random chance, and negative deviation from zero indicates a decreased observed frequency of the sister-lipotype combination compared with random chance. P values were calculated using the bootstrap pairwise t test (**Methods 4.3.9**) **(F)** Probability of a lipotype state transition occurring in a cell over a 21-hour time period as estimated using CELLMA (**Methods 4.3.1-4.4.6**). Probabilities are located at the corresponding arrow tails. **(G)** Markov model–simulated evolution of a pure ChTxB+ (left) or pure ShTxB1a+/2e+ (right) cell population over 7 days. **(H)** Line plots displaying the evolution of the state predictability of a cell (or its progeny) after a certain time from an original state measurement. Differently colored tracks correspond to a different original lipotype measurement (t = 0). Kullback-Leibler

divergence is evaluated between the probability distribution vector obtained using the Markov transition matrix and the steady-state probability distribution (i.e., the best uninformed guess).

4.3.3. CELLMA infers transition probabilities and lipotype stabilities

Considering the toxin-staining patterns of lineage-related cells, we modeled the dynamics of lipotype state transitions with a cell-state transition estimation by lineage leaf-state Markov analysis (CELLMA) (**Methods 4.5.1-4.5.6**). CELLMA calculates a Markov transition probability matrix among all combinations of lipotypes. This transition matrix offers insight into the interconversion fluxes and stability of the different pre-defined lipid states.

We applied CELLMA to our data and found that ShTxB1a$^+$/2e$^+$, ChTxB$^+$, and triple$^+$ are the most stable states, with a 37%, 51%, and 68% probability of converting into a different lipotype during a single cell replication cycle (21 hours), respectively. On the other hand, ShTxB1a$^+$ and ShTxB2e$^+$ states showed a greater likelihood (95 and 80% during a single replication cycle, respectively) of converting into the ShTxB1a$^+$/2e$^+$ or into “other” lipid configurations (**Fig. 4.2F and 4.1SE-F**). This suggests the ShTxB1a/2e$^+$, ChTxB+, and Triple$^+$ states act as major attractor states that can slowly interconvert and propagate across cell generations, whereas the ShTxB1a$^+$ and ShTxB2e$^+$ states are more transient.

Furthermore, our model predicted that lipotype homogeneous populations, composed of cells all belonging to the same lipid category, would revert slowly (i.e., up to 7 days for the slowest case of ChTxB+) to a lipotype heterogeneous steady state (**Fig. 4.2G-H and S4.1G**). In agreement with this hypothesis, when we selected ShTxB1a$^+$ or ChTxB$^+$ cells by fluorescence-activated cell sorting (FACS) and kept them in culture for 10 days, the cultures reverted to heterogeneous cell populations with lipid-state compositions similar to the cultures from which they were originally selected (**Fig. S4.1H**). This aligns with our conclusion that dHFs exist in metastable sphingolipid configurations at homeostasis.

4.4. Discussion

In this work, we introduce the concept of lipotypes to describe recurring metabolic states in dHFs. We propose a computational approach called CELLMA to infer the probability matrix representing lipid-state transitions. CELLMA harnesses time-lapse microscopy to reconstruct cell lineages and endpoint measurements of lipid state through the proxy of toxin stainings at a single-cell spatial resolution (**Fig. 4.2A**). Our model offers a unique tool to study recurring, acyclical cell state transitions in biological systems.

Interestingly, we discovered that lipotypes tend to have varying degrees of stability (**Fig. 4.2F**). This highlights an intriguing aspect of cellular identity, which is that even in steady state biological settings, not all cell states are equal in duration. It is still unclear what signals

are necessary to drive cell state transitions in dHF cultures and whether disruptions to those signals are commonplace. In the future, CELLMA could be extended to model lipotypes determined by a combination of individual lipid measurements, rather than from toxin stainings directly (**Fig. A4.1**). This would offer insight into the potential metabolic mechanisms driving lipotype transitions. Furthermore, this would offer a greater diversity of features with which to define lipotype states, since some lipid classes (i.e., HexCer) cannot be isolated by toxin staining.

Although it has been suggested before that lipids contribute to the definition of cell state (Frechin et al., 2015; Kramer et al., 2022), the findings presented here and in the complete published study (Capolupo, Khven, et al., 2022) further confirm that lipid-defined states align well with phenotypic states defined by other cellular features. dHFs appear to maintain enough plasticity to reach all possible lipotypes, or at least those represented by toxin stainings. These conclusions support the hypothesis that lipid remodeling can have long-term effects on cellular identity. Lipids could define more complex cell states and drive cell state transitions in other biological settings too, particularly during differentiation and embryonic development (Bhaduri et al., 2021; D'Angelo & La Manno, 2023; Levental & Lyman, 2023; D. Russo, Della Ragione, et al., 2018; Yoon et al., 2022). In these settings, the assumption made by CELLMA that all lipotypes are present in constant proportions may no longer hold, and changes to the model's formulation to allow for multiple transition matrices would be necessary. Ultimately, the ability to measure lipid abundance at a single-cell resolution opens new opportunities for computational models to infer cell state transitions as defined by cellular metabolic activity.

4.5. Methods

4.5.1. Introduction of the model and data

Here, we formulate a model that describes the behavior of a pool of cells that can transition between discrete, mutually exclusive "states" and divide, generating two daughter cells. Specifically, we consider the scenario where the relative abundances of cells in different states do not vary significantly over time, while cells continuously undergo transitions (i.e., the cell pool is at steady-state).

The model we present is designed in such a way to allow the estimation of the rates of cell state transitions from incomplete recordings. In particular, we are interested in the situation where one can measure the time-resolved division history of hundreds of cells in a pool, but not their evolving state. If the time-resolved information on the cell state were available, the estimation of the cell-state transition rates would be direct and rather trivial.

Instead, the limitation that requires computationally maneuvering is that cell state cannot be probed for its state “live” and can only be recorded at an endpoint after measuring the division history.

Pragmatically, measurements of this sort could be the result of slightly different experimental set-ups. We consider a case where live imaging microscopy has been used to record the division history and lipid anti-toxin stainings to assess the endpoint state of the cells. Live imaging recordings, appropriately processed by cell segmentation and tracking, provide both kinship relationships between cells and the time of division events, a variable important for the modeling framework presented here. The information of cell staining can be summarized in a modest number of discretized cell states, among which we want to estimate the rate of transition. Here we assume five unique lipid configurations, as determined by anti-toxin staining at the time lapse end point: ChTxB+, ShTxB1+, ShTxB2+, ShTxB1+ShTxB2+ (double), ShTxB1+ShTxB2+ChTxB+ (triple). We finally consider an additional sixth lipid configuration (Other), which corresponds to states that cannot be uniquely resolved by toxin stainings alone.

Furthermore, in the biological context considered in the present study, we have no ways of measuring “live” casual factors or covariates correlating with the transition choice of an individual cell. Thus, for the purpose of our model, the state transition occurring in a particular cell can be modeled as a stochastic process, yet with fixed rates. We also note that there is no biological reason to assume that the past state of a cell should affect the likelihood of the cell to transition from its current to a new, future state. Also, we have no biological evidence suggesting the system might not be ergodic.

These considerations naturally lead to the choice of a Markov chain to model the transition process. Finally, we assume that cell division itself does not induce an immediate state transition. In biological terms, this assumption means that we do not allow for asymmetric segregation of the lipids during cell division. However, we note that if this were not true, our fit should accommodate that occurrence by skewing the transition rates to recapitulate a similar behavior.

In summary, the model we propose considers a discrete-time Markov chain evolution of the cell state in each cell of the pool. Transitions are not affected by cell divisions. The model postulates that two daughter cells will inherit the same cell state from their mother at the moment of division. Only then will their states start drifting apart, as they will be more likely to transition to other states according to the same transition probability.

We mathematically formulate our model as a maximum likelihood estimation of the transition matrix between cell states, and we anticipate that it will formally show that one can

use the knowledge of the time at which sister-cell drift started together with the recording of the cell's final state to estimate the transition matrix of the Markov chain.

4.5.2. CELLMA model formulation

Let us indicate with $\mathbf{c}_i \in [0,1]^n$: $\mathbf{c}_i^T \cdot \mathbb{1} = 1$ the vector containing the probabilities that cell i is in each of n possible states. Upon experimental measurement, we become certain that a cell is in state s and therefore $\mathbf{c}_i = \mathbf{e}_s$, where the notation $\mathbf{e}_s$ indicates a unit vector with its s-th element equal to one and the other entries equal zero. We also define, for convenience, $\mathcal{S} = \{\mathbf{e}_s \forall s \in \mathbb{N}: 1 \leq s \leq n\}$.

We consider a discrete-time setting and postulate that for a single time step, the state of a cell changes following a memoryless Markov process, defined by the transition matrix $\mathbf{\Theta} \in [0,1]^{n \times n}$, where each entry Θ_{ij} is the transition probability from state j to state i. Thus, the updated probability vector after a series of k time steps can be easily obtained by left-multiplying the vector with the matrix. For example, the updated probability vector of a cell i after $k \in \mathbb{N}$ time steps can be written as:

$$\mathbf{c}_{i|t+k} = \mathbf{\Theta}^k \cdot \mathbf{c}_{i|t} \qquad (\mathbf{1})$$

Where $\mathbf{\Theta}^k$ indicates the matrix power and the notation $_{|t}$ is used to specify that the state that a particular variable is considered at a particular time point t. We note that $\mathbf{\Theta}$ is a stochastic matrix, that is $\mathbf{\Theta}^T \cdot \mathbb{1} = \mathbb{1}$.

From the above, it follows that the probability of a cell to be in state s starting from a state probability vector $\mathbf{c}_i$ can be written as:

$$P(\mathbf{c}_{i|t+k} = \mathbf{e}_s \mid \mathbf{c}_{i|t}, \mathbf{\Theta}) = \mathbf{e}_s^T \cdot \mathbf{\Theta}^k \cdot \mathbf{c}_{i|t} \qquad (\mathbf{2})$$

Let us now consider the notation for data. For each cell at the end of our division tracking experiment, we readout a cell state. Consistently with the formulation defined above, we represent this state using a unit vector. The cell state measured for cell i will be, thus, indicated as $\mathbf{m}_i \in \mathcal{S}$. Conversely, the state of each parent cell, which we cannot measure, has to be considered as a latent variable. For an easier discrimination between measurable and latent cell states, we indicate those two sets of variables with $\mathbf{c}$ and γ respectively. Furthermore, we indicate the sister relationship between two cells i and j writing $j = \alpha(i)$ and the mother-daughter relation between cell k and i writing $k = \mu(i)$ and conversely $i = \mu^{-1}(k) = \hbar(k)$. Finally, we indicate a set of relatives as a “clade” and distinguish the set of

measured states in the r th clade with $\mathcal{K}_r$ and the set of latent (e.g., parent) states in that clade with $\mathcal{Q}_r$. Considering a cell i we then have $\mathcal{K}_r = \kappa(i)$ and $\mathcal{Q}_r = \rho(i)$, so that we can write expressions such as $\alpha(\mu(i)) \in \mathcal{Q}_r$.

From now on we will omit the notation $_{|t}$ and, instead, indicate with $\mathbf{c}_i$ the state of the i-th cell at the time of the final measurement and with γ_j the state of the j-th latent cell at the time when it divides generating the two daughter cells. We note that all the equations below can be written as a function of those variables only and of the number of time-separating pairs of related cells i and j that we indicate with $\chi_{ji} \in \mathbb{N}$.

4.5.3. Three assumptions of cell lipid configuration heritability

The aim of CELLMA is to produce a Maximum Likelihood estimate $\hat{\boldsymbol{\Theta}}$ of the Markov transition matrix of the state transition process from the set of data $\mathcal{M} = \{\mathbf{m}_i\}$. In order to achieve this, we need to consider the following three facts and assumptions. Firstly, sister cell states are not independent because at time of their generation have inherited the same cell type from their mother cell.

$$P_{\boldsymbol{\Theta}}\big(\mathbf{c}_i = \mathbf{e}_z, \mathbf{c}_j = \mathbf{e}_q\big) \neq P_{\boldsymbol{\Theta}}(\mathbf{c}_i = \mathbf{e}_z) \cdot P_{\boldsymbol{\Theta}}\big(\mathbf{c}_j = \mathbf{e}_q\big)\ \forall (i,j)\colon j = \alpha(i) \qquad (\mathbf{3})$$

Instead, sister cells are conditionally independent given the mother cell state at the time of division and the Markov transition matrix, as the two cells in an identical state after division transition to new state independently following a Markov process.

$$P_{\boldsymbol{\Theta}}\big(\mathbf{c}_i = \mathbf{e}_z, \mathbf{c}_j = \mathbf{e}_q \mid \boldsymbol{\gamma}_b\big) = P_{\boldsymbol{\Theta}}(\mathbf{c}_i = \mathbf{e}_z \mid \boldsymbol{\gamma}_b) \cdot P_{\boldsymbol{\Theta}}\big(\mathbf{c}_j = \mathbf{e}_q \mid \boldsymbol{\gamma}_b\big) \forall (i,j,b)\colon j = \alpha(i) \& b = \mu(i) \qquad (\mathbf{4})$$

The state of cells for which we do not observe any evidence of kinship during the time-lapse imaging and are, thus, from different "clades" are considered independent beside having to obey the same probabilistic process determined by the Markov matrix.

$$P_{\boldsymbol{\Theta}}\big(\mathbf{c}_i = \mathbf{e}_z, \mathbf{c}_j = \mathbf{e}_q\big) = P_{\boldsymbol{\Theta}}(\mathbf{c}_i = \mathbf{e}_z) \cdot P_{\boldsymbol{\Theta}}\big(\mathbf{c}_j = \mathbf{e}_q\big)\ \forall (i,j)\colon j \notin \kappa(i) \qquad (\mathbf{5})$$

4.5.4. Model simplification with maximum likelihood estimation

We have now all the ingredients to simplify the expression of the Maximum likelihood problem to estimate $\boldsymbol{\Theta}$.

$$\begin{aligned} \underset{\boldsymbol{\Theta}}{\text{maximize}} \quad & \mathcal{L}(\boldsymbol{\Theta} \mid \{\mathbf{m}_i\}) \\ \text{s.t.} \quad & \boldsymbol{\Theta}^T \cdot \mathbb{1} = \mathbb{1} \\ & \boldsymbol{\Theta} \succcurlyeq 0 \end{aligned} \qquad (\mathbf{6})$$

With $\mathcal{L}(\boldsymbol{\Theta} \mid \{\mathbf{m}_i\}) = P_{\boldsymbol{\Theta}}(\{\mathbf{c}_i = \mathbf{m}_i\})$. First, we consider the assumption expressed equation (5) to factor the likelihood function by clades as follows:

$$P_{\boldsymbol{\Theta}}(\{\mathbf{c}_i = \mathbf{m}_i\}) = \prod_{r=1}^{N_K} P_{\boldsymbol{\Theta}}(\{\mathbf{c}_i = \mathbf{m}_i \forall i \in \mathcal{K}_r\}) \qquad (\mathbf{7})$$

where N_K is the number of clades.

Now we note that we can compute the likelihood for a clade considering the joint probability of the data and the latent variables of a clade, and by marginalizing over all the possible combination of states. More specifically we have:

$$\prod_{r=1}^{N_K} P_{\boldsymbol{\Theta}}(\{\mathbf{c}_i = \mathbf{m}_i \forall i \in \mathcal{K}_r\})$$

$$= \prod_{r=1}^{N_K} \sum_{l=1}^{n} \sum_{k=1}^{n} \dots \sum_{q=1}^{n} P_{\boldsymbol{\Theta}}\left(\{\mathbf{c}_i = \mathbf{m}_i \forall i \in \mathcal{K}_r\}, \gamma_a = \mathbf{e}_l, \gamma_b = \mathbf{e}_k, \dots \gamma_q = \mathbf{e}_q\right) \qquad (\mathbf{8})$$

Where only the latent variables in that clade are considered (e.g., $\{a, b, \dots, z\} = \mathcal{Q}_r$). We can write the equation (8) more compactly considering two compounded variables, one indicating the possible combination of latent states $\mathbf{E} = \left(\mathbf{e}_l, \mathbf{e}_k, \dots, \mathbf{e}_q\right) \in \mathcal{S}^{|\mathcal{Q}_r|}$ that will be indexed for convenience as follows $\mathbf{E} = (\varepsilon_a, \varepsilon_b, \dots \varepsilon_z) : \{a, b, \dots, z\} = \mathcal{Q}_r$ and the other corresponding to all stacked latent variables of a clade $\boldsymbol{\Gamma}_r = (\gamma_a, \gamma_b, \dots \gamma_z) : \{a, b, \dots, z\} = \mathcal{Q}_r$. Note that $\mathcal{S}^{|\mathcal{Q}_r|}$ indicates a Cartesian power and $|\cdot|$ indicates the carnality of a set.

$$P_{\boldsymbol{\Theta}}(\{\mathbf{c}_i = \mathbf{m}_i \forall i : i \in \mathcal{K}_r\}) = \sum_{\mathbf{E} \in \mathcal{S}^{|\mathcal{Q}_r|}} P_{\boldsymbol{\Theta}}(\{\mathbf{c}_i = \mathbf{m}_i \forall i : i \in \mathcal{K}_r\}, \boldsymbol{\Gamma}_r = \mathbf{E}) \qquad (\mathbf{9})$$

We note that by relying on the conditional independence of daughter cell state given the mother (e.g., equation (4)), we can factorize the term inside the sum as the product of probabilities for each cell, summed over all possible states $\mathcal{S}$ for the parent cells, which are latent variables.

$$P_{\boldsymbol{\Theta}}(\{\mathbf{c}_i = \mathbf{m}_i i \in \mathcal{K}_r\}) = \sum_{\mathbf{E} \in \mathcal{S}^{|\mathcal{Q}_r|}} \prod_{\forall i \in \mathcal{K}_r} P_{\boldsymbol{\Theta}}(\mathbf{c}_i = \mathbf{m}_i, \boldsymbol{\Gamma}_r = \mathbf{E}) \qquad (\mathbf{10})$$

Despite the appearance, the computation of equation (10) is less daunting than it may seem at first sight as the expression also factorizes in the "mother-daughter" terms shown

below (either latent-measured or latent-latent). Those terms have an analogous form of the expression in equation (2) and allow us to explicit the function of the only unknown quantity $\boldsymbol{\Theta}$, resulting in a functional form that is the product of different terms with:

$$\begin{aligned}
&\sum_{\mathbf{E}\in\mathcal{S}^{|\mathcal{Q}_r|}} \prod_{i\in\mathcal{K}_r} P_{\boldsymbol{\Theta}}(\mathbf{c}_i = \mathbf{m}_i, \boldsymbol{\Gamma}_r = \mathbf{E}) = \\
&= \sum_{\mathbf{E}\in\mathcal{S}^{|\mathcal{Q}_r|}} \prod_{i\in\mathcal{K}_r} P_{\boldsymbol{\Theta}}\big(\mathbf{c}_i = \mathbf{m}_i, \gamma_{\mu(i)} = \varepsilon_{\mu(i)}\big) \prod_{u\in\mathcal{Q}_r} P_{\boldsymbol{\Theta}}\big(\gamma_u = \varepsilon_u, \gamma_{\mu(u)} = \varepsilon_{\mu(u)}\big) = \\
&= \sum_{\mathbf{E}\in\mathcal{S}^{|\mathcal{Q}_r|}} \prod_{i\in\mathcal{K}_r} \mathbf{m}_i^T \cdot \boldsymbol{\Theta}^{\chi_{i,\mu(i)}} \cdot \varepsilon_{\mu(i)} \prod_{u\in\mathcal{Q}_r} \varepsilon_u^T \cdot \boldsymbol{\Theta}^{\chi_{u,\mu(u)}} \cdot \varepsilon_{\mu(u)}
\end{aligned} \qquad (\mathbf{11})$$

Where we have indicated with $\mathbf{E}_{\mu(i)}$ the unitary vector $\mathbf{e}_s$ extracted at the $\mu(i)$-th row of E.

We can finally combine everything we have learned in a single expression of the likelihood function:

$$\mathcal{L}(\boldsymbol{\Theta} \mid \{\mathbf{m}_i\}) = \prod_{r=1}^{N_K} \sum_{\mathbf{E}\in\mathcal{S}^{|\mathcal{Q}_r|}} \prod_{i\in\mathcal{K}_r} P_{\boldsymbol{\Theta}}(\mathbf{c}_i = \mathbf{m}_i, \boldsymbol{\Gamma}_r = \mathbf{E}) \qquad (\mathbf{12})$$

Explicating all the terms as a function of $\boldsymbol{\Theta}$ and the data, and listing all the constraints, we arrive to the final form of the CELLMA optimization problem:

$$\begin{aligned}
&\underset{\boldsymbol{\Theta}}{\text{maximize}} \prod_{r=1}^{N_K} \sum_{\mathbf{E}\in\mathcal{S}^{|\mathcal{Q}_r|}} \prod_{\forall i: i\in\mathcal{K}_r} \mathbf{m}_i^T \cdot \boldsymbol{\Theta}^{\chi_{i,\mu(i)}} \cdot \boldsymbol{\varepsilon}_{\mu(i)} \prod_{\forall u: u\in\mathcal{Q}_r} \boldsymbol{\varepsilon}_u^T \cdot \boldsymbol{\Theta}^{\chi_{u,\mu(u)}} \cdot \boldsymbol{\varepsilon}_{\mu(u)} \\
&\text{s.t. } \boldsymbol{\Theta}^T \cdot \mathbb{1} = \mathbb{1} \\
&\quad\ \boldsymbol{\Theta} \geqslant 0
\end{aligned} \qquad (\mathbf{13})$$

4.5.5. Implementation practicalities

In practice, at the implementation level we actually minimize the following negative log-likelihood function:

$$\begin{aligned}
&\underset{\boldsymbol{\Theta}}{\text{minimize}} -\log\, \mathcal{L}(\boldsymbol{\Theta} \mid \{\mathbf{m}_i\}) \\
&\underset{\boldsymbol{\Theta}}{\text{minimize}} -\sum_{r=1}^{N_K} \log \sum_{\mathbf{E}\in\mathcal{S}_{|\mathcal{Q}_r|}} \prod_{i\in\mathcal{K}_r} P_{\boldsymbol{\Theta}}(\mathbf{c}_i = \mathbf{m}_i, \boldsymbol{\Gamma}_r = \mathbf{E})
\end{aligned}$$

For numerical stability we use log-probabilities whenever possible to avoid numerical cancellation problems. In particular we solve the problem in this form:

$$\underset{\boldsymbol{\Theta}}{\text{minimize}} - \sum_{r=1}^{N_K} \log \sum_{\mathbf{E} \in \mathcal{S}^{|\mathcal{Q}_r|}} \exp \left(\sum_{i \in \mathcal{K}_r} \log \left(\mathbf{m}_i^T \cdot \boldsymbol{\Theta}^{\chi_{i,\mu(i)}} \cdot \boldsymbol{\varepsilon}_{\mu(i)} \right) + \sum_{u \in \mathcal{Q}_r} \log \left(\varepsilon_u^T \cdot \boldsymbol{\Theta}^{\chi_{u,\mu(u)}} \cdot \boldsymbol{\varepsilon}_{\mu(u)} \right) \right)$$

$$\text{s.t. } \boldsymbol{\Theta}^T \cdot \mathbb{1} = \mathbb{1}$$

$$\boldsymbol{\Theta} \succcurlyeq 0$$

We solve the above problem using a standard interior point solver. We note that the gradient of the function above is not tractable analytically, nonetheless we can compute it empirically at each iteration of the solver. Clearly, this adds to the computational cost quadratically with respect to the number of states. More influential on time complexity is, however, is the Cartesian product that brings in a factorial contribution. However, we notice that more distant ancestors are progressively less crucial to the estimation, and therefore we can partition big clades by dropping all lineage relation deeper than three layers without losing much information. Overall, for several hundreds of cells and a small number of states (e.g., ≤ 10) the method converges in just few minutes.

4.5.6. Exploiting the hierarchical structure to economize the computation

We note that when we made a step from (8) to (9) we ignored the structure of the tree seeking only a general formulation. However, by doing so, at the level of implementation we end up repeating computations where the structure of the tree might allow a much more compact computation. We will show here that, with a bit of bookkeeping, it is possible to reduce significantly the complexity of the likelihood function evaluation.

Let's consider the subclade defined taking an arbitrary internal node γ_i (let's call it the subcalde root) and its progeny. More formally $\mathcal{W}_i = \{\forall j : \exists n \mu^n(j) = i\}$. We can define a conditional likelihood of such a subclade, indicated with $\mathcal{L}_\Theta^{i|q}$, as the probability of the data observed downstream of the subclade root γ_i, conditional on it having state q. In other words:

$$\mathcal{L}_{\boldsymbol{\Theta}}^{i|q} = P_{\boldsymbol{\Theta}}\left(\{\mathbf{c}_j = \mathbf{m}_j \forall j \in \mathcal{W}_i \cap \mathcal{K}_r\} \mid \gamma_i = \mathbf{e}_q\right)$$

that can be written explicitly as a marginalization of the following joint probability distribution:

$$\mathcal{L}_\Theta^{i|q} = \sum_{l=1}^{n} \sum_{k=1}^{n} \dots \sum_{z=1}^{n} P_{\boldsymbol{\Theta}}\left(\{\mathbf{c}_j = \mathbf{m}_j \forall j \in \mathcal{W}_i \cap \mathcal{K}_r\}, \gamma_\delta = \mathbf{e}_l, \gamma_o = \mathbf{e}_k, \dots, \gamma_\omega = \mathbf{e}_z \mid \gamma_i = \mathbf{e}_q\right)$$

$$\text{with } \{\delta, o, \dots, \omega\} \doteq \mathcal{W}_i \cap \mathcal{Q}_r$$

Let's now consider a bisection of the set $\{\delta, o, \dots, \omega\}$ in two sets of nodes (and corresponding latent variables) each corresponding of the nodes downstream one of the two children of the node i (note, the corresponding latent variables are $\gamma_{\hbar(i)_1}$ and $\gamma_{\hbar(i)_2}$). We can write these sets compactly as $\{\hbar(i)_1, \delta_1, \dots, \omega_1\} \doteq \mathcal{W}_{\hbar(i)_1} \cap \mathcal{Q}_r$ and $\{\hbar(i)_2, \delta_2, \dots, \omega_2\} \doteq \mathcal{W}_{\hbar(i)_2} \cap \mathcal{Q}_r$. Noticing then that the latent variables from one of these sets are independent to the one of the other conditional on γ_i we can factorize the above:

$$\mathcal{L}_{\boldsymbol{\Theta}}^{i|q} = \left(\sum_{l=1}^{n}\sum_{k=1}^{n}\dots\sum_{z=1}^{n} P_{\boldsymbol{\Theta}}\left(\left\{\mathbf{c}_j = \mathbf{m}_j \forall j \in \mathcal{W}_i \cap \mathcal{K}_r\right\}, \boldsymbol{\gamma}_{\hbar(i)_1} = \mathbf{e}_l, \boldsymbol{\gamma}_{\delta_1} = \mathbf{e}_k, \dots, \boldsymbol{\gamma}_{\omega_1} = \mathbf{e}_z \mid \gamma_i = \mathbf{e}_q\right)\right) \cdot \left(\sum_{p=1}^{n}\sum_{t=1}^{n}\dots\sum_{r=1}^{n} P_{\boldsymbol{\Theta}}\left(\left\{\mathbf{c}_j = \mathbf{m}_j \forall j \in \mathcal{W}_i \cap \mathcal{K}_r\right\}, \gamma_{\hbar(i)_2} = \mathbf{e}_p, \gamma_{\delta_2} = \mathbf{e}_t, \dots, \gamma_{\omega_2} = \mathbf{e}_r \mid \boldsymbol{\gamma}_i = \mathbf{e}_q\right)\right)$$

And now, using the definition of conditional probability and exploiting the fact that the cell state of a cell is independent from the state of the grandmother (or more remote progenitors) conditional on the state mother we can write:

$$\mathcal{L}_{\boldsymbol{\Theta}}^{i|q} = \left(\sum_{l=1}^{n}\sum_{k=1}^{n}\dots\sum_{z=1}^{n} P_{\boldsymbol{\Theta}}(\gamma_{\hbar(i)_1} = \mathbf{e}_l | \gamma_i = \mathbf{e}_q) \cdot P_{\boldsymbol{\Theta}}\left(\left\{\mathbf{c}_j = \mathbf{m}_j \forall j \in \mathcal{W}_{\hbar(i)_1} \cap \mathcal{K}_r\right\},\right.\right.$$
$$\left.\left.\gamma_{\delta_1} = \mathbf{e}_k, \dots, \boldsymbol{\gamma}_{\omega_1} = \mathbf{e}_z \mid \gamma_{\hbar(i)_1} = \mathbf{e}_l\right)\right) \cdot \left(\sum_{p=1}^{n}\sum_{t=1}^{n}\dots\sum_{r=1}^{n} P_{\boldsymbol{\Theta}}\left(\gamma_{\hbar(i)_2} = \mathbf{e}_p \mid \boldsymbol{\gamma}_i = \mathbf{e}_q\right) \cdot\right.$$
$$\left.P_{\boldsymbol{\Theta}}\left(\left\{\mathbf{c}_j = \mathbf{m}_j \forall j \in \mathcal{W}_{\hbar(i)_2} \cap \mathcal{K}_r\right\}, \gamma_{\delta_2} = \mathbf{e}_t, \dots, \gamma_{\omega_2} = \mathbf{e}_r \mid \gamma_{\hbar(i)_2} = \mathbf{e}_p\right)\right) =$$
$$\left(\sum_{l=1}^{n} P_{\boldsymbol{\Theta}}(\gamma_{\hbar(i)_1} = \mathbf{e}_l | \boldsymbol{\gamma}_i = \mathbf{e}_q) \cdot \sum_{k=1}^{n}\dots\sum_{z=1}^{n} P_{\boldsymbol{\Theta}}\left(\left\{\mathbf{c}_j = \mathbf{m}_j \forall j \in \mathcal{W}_{\hbar(i)_1} \cap \mathcal{K}_r\right\},\right.\right.$$
$$\left.\left.\gamma_{\delta_1} = \mathbf{e}_k, \dots, \boldsymbol{\gamma}_{\omega_1} = \mathbf{e}_z \mid \gamma_{\hbar(i)_1} = \mathbf{e}_l\right)\right) \cdot \left(\sum_{p=1}^{n} P_{\boldsymbol{\Theta}}\left(\gamma_{\hbar(i)_2} = \mathbf{e}_p \mid \boldsymbol{\gamma}_i = \mathbf{e}_q\right) \cdot\right.$$
$$\left.\sum_{t=1}^{n}\dots\sum_{r=1}^{n} P_{\boldsymbol{\Theta}}\left(\left\{\mathbf{c}_j = \mathbf{m}_j \forall j \in \mathcal{W}_{\hbar(i)_2} \cap \mathcal{K}_r\right\}, \boldsymbol{\gamma}_{\delta_2} = \mathbf{e}_t, \dots, \boldsymbol{\gamma}_{\omega_2} = \mathbf{e}_r \mid \boldsymbol{\gamma}_{\hbar(i)_2} = \mathbf{e}_p\right)\right)$$

That can be compactly written in a recursive relation, noticing that the second term in each factor has the form of:

$$\mathcal{L}_{\boldsymbol{\Theta}}^{i|q} = \left(\sum_{l=1}^{n} P_{\boldsymbol{\Theta}}\left(\gamma_{\hbar(i)_1} = \mathbf{e}_l \mid \gamma_i = \mathbf{e}_q\right) \cdot \mathcal{L}_{\boldsymbol{\Theta}}^{\hbar(i)_1 l}\right) \cdot \left(\sum_{p=1}^{n} P_{\boldsymbol{\Theta}}\left(\gamma_{\hbar(i)_2} = \mathbf{e}_p \mid \gamma_i = \mathbf{e}_q\right) \cdot \mathcal{L}_{\boldsymbol{\Theta}}^{\hbar(i)_2 | p}\right) =$$

$$\left(\sum_{l=1}^{n} \mathbf{e}_l^T \boldsymbol{\Theta}^{\chi_{\hbar(i)_1, i}} \mathbf{e}_q \cdot \mathcal{L}_{\boldsymbol{\Theta}}^{\hbar(i)_1 l}\right) \cdot \left(\sum_{p=1}^{n} \mathbf{e}_p^T \boldsymbol{\Theta}^{\chi_{\hbar(i)_2, i}} \mathbf{e}_q \cdot \mathcal{L}_{\boldsymbol{\Theta}}^{\hbar(i)_2 | p}\right)$$

We can write it in a more compact form introducing the vector $\ell_{\Theta}^{i} = \left[\mathcal{L}_{\Theta}^{i|1}, \mathcal{L}_{\Theta}^{i|2}, \ldots, \mathcal{L}_{\Theta}^{i|n}\right]$.

$$\mathcal{L}_{\Theta}^{i|q} = \left(\ell_{\Theta}^{\hbar(i)_1^T} \cdot \boldsymbol{\Theta}^{\chi_{\hbar(i)_1, i}} \cdot \mathbf{e}_q\right)\left(\ell_{\Theta}^{\hbar(i)_2^T} \cdot \boldsymbol{\Theta}^{\chi_{\hbar(i)_2, i}} \cdot \mathbf{e}_q\right)$$

And even more compactly:

$$\boldsymbol{\ell}_{\boldsymbol{\Theta}}^{i\ T} = \boldsymbol{\ell}_{\boldsymbol{\Theta}}^{\hbar(i)_1 T} \cdot \boldsymbol{\Theta}^{\chi_{\hbar(i)_1, i}} \odot \boldsymbol{\ell}_{\boldsymbol{\Theta}}^{\hbar(i)_2 T} \cdot \boldsymbol{\Theta}^{\chi_{\hbar(i)2}, i}$$

To express eq. (7) in terms of this recursion we need to specify how to deal with root node of clade r that we indicate $\pi(r)$. Similarly of what is done in the naive computation of the likelihood, we consider a prior ψ that we typically will set to the sample distribution of states. Note that this corresponds to a steady state assumption. However, a more informative prior (e.g., only one "stem cell" population being present at the origin) can be used as well. So, we can write the final likelihood $\mathcal{L}(\boldsymbol{\Theta} \mid \{\mathbf{m}_i\})$ as

$$P_{\boldsymbol{\Theta}}(\{\mathbf{c}_i = \mathbf{m}_i\}) = \prod_{r=1}^{N_K} \ell_{\boldsymbol{\Theta}}^{\pi(r)^T} \boldsymbol{\Theta}^{\chi, \pi(r)} \boldsymbol{\psi}$$

or, depending on what the prior is intended to represent just

$$P_{\boldsymbol{\Theta}}(\{\mathbf{c}_i = \mathbf{m}_i\}) = \prod_{r=1}^{N_K} \ell_{\boldsymbol{\Theta}}^{\pi(r)^T} \boldsymbol{\psi}$$

4.5.7. Lipid determination from toxin stainings

After segmenting each cell, we determined its lipotype class by considering a binarization to each toxin signal i.e., designating each cell as either “positive” or “negative”. To achieve this, we first normalized the signal intensity to control for cell size and adjust the log-response of the fluorescent intensity signal. Specifically, we used the normalized values $Z_{ic} = log(5\ R_{ic} / max_c(R_{ic}))$, where $R_{ic} = I_{ic}/\Sigma_z I_{zc}$ and I_{ic} indicates the signal of the toxin i in cell c. We then thresholded Z_{ic} to obtain the binarization. To accommodate for the potential experimental fluctuation of the total maximum intensity of signal, we used sample-relative

thresholds determined identically for each of the wild type unperturbed dHF stainings. A cell was considered ChTxB+ if the level of ChTx was higher than the 35th percentile, ShTxB1a+ if its ShTxB1a signal exceeded the 40th percentile, ShTxB2e+ if its ShTxB2e signal was higher than the 25th percentile.

4.5.8. Time-lapse imaging coupled with endpoint staining

dHFs were seed the day before on glass bottom 3 well chamber slides (IBIDI) in complete media to reach the roughly 30% confluence at the start of the time-lapse. Cells were analyzed on a PerkinElmer Operetta microscope under controlled temperature and CO_2 and followed for 42 hours. An image was acquired every 20 minutes in brightfield and digital phase contrast (DPC) with a 10x (0.35 NA) air objective, binning of 2 and the speckle scale set to 0 under non-saturated conditions. After 42h, cells were fixed with 4% PFA and processed for toxin staining. The same areas acquired by brightfield and DPC microscopy were analyzed with Leica SP8 with 20x air objective (0.8 NA) as described above.

4.5.9. Cell state transition estimation from time-resolved lineages

The lineage information extracted from the time-lapse imaging and the cell state from the endpoint staining were used to perform a sister state frequency analysis and to fit CELLMA. First, we segmented the time-lapse images using a custom Cellpose model trained using manually annotated dHFs DPC micrographs (Capolupo, 2022; Capolupo, Burri, et al., 2022). All the segmentation hypotheses were, then, combined to allow for cell tracking. Cell tracking was performed with TrackMate v7.2 using the “LAP tracker” algorithm and allowing track segment splitting and gap closing. The following parameters were used: frame-to-frame linking max distance: 22 µm, track segment gap closing max distance: 20 µm and max frame gap: 6 frames, track segment splitting max distance 25 µm. Tracking information was exported from TrackMate and analyzed using a custom Python script. The final lineage tree analyzed was obtained by pruning the raw tracks in the following way: we considered only the tracks that reached the last frame, where at least one splitting event was recorded and that had a cumulative length of at least 180 µm. Cumulatively, we analyzed a total of 1,516 leaf cells grouped into 591 clades (e.g., cells deriving from an initial mother cell). No manual curation of the tracks was performed. The last frame of the time-lapse was manually registered to the toxin staining image and the segmentation of the last frame was used to quantify the toxin levels of each cell. Toxin signal quantifications were normalized and cell state determined as described in section 4.3.6.

Sister state pair frequency was measured by simply counting the occurrences of all the possible pairs of sister cell states. A one-sided Monte Carlo test was performed for each possible state pair in the attempt to reject the null hypothesis that the occurrence of that state pair in sister cells is compatible with a random state assignment. More precisely, we generated 100,000 realizations of the null model by permuting the cell labels while leaving the lineage tree unaltered. To estimate the tendency of each lipotype to transition into another, provided only the toxin-stained endpoint readout and the time-resolved lineage traces, we then applied CELLMA as formulated in sections 4.3.1-4.3.4.

CELLMA minimizes the negative log-likelihood defined with respect to the Markov matrix $\boldsymbol{\theta}$, subjected to both linear inequality (i.e., non-negativity) and equality (i.e., to ensure that θ is a stochastic matrix) constraints. We used an interior point routine developed for large scale programs (scipy v1.71 implementation function "minimize('trust-const')"). The time step for the transition encoded by θ was selected to be 100 min. Transition matrices corresponding to shorter steps were obtained by evaluating the matrix power. Since this optimization problem is not convex, we run the optimization 12 times starting with random guesses for 50 iterations, the solution candidate that minimizes the negative log-likelihood is then further refined for 500 iterations or if convergence is reached. The random starting guesses are generated renormalizing $Mij\sim\ \delta ij$ + 15 $Beta$(0.3, 0.5) (where δij is the Kroneker delta) to a stochastic matrix. To allow efficient CELLMA model fitting, only latent variables two generations above the leaf cells were considered deeper clades were split. Time-resolved simulations of cell state evolution were performed by repeatedly left-multiplying an initial probability distribution vector by θ. As a proxy for the future state predictability of a lipotype after a time Δt passed since its state $p0$ was measured, we use the Kullback-Leibler (KL) divergence between the expected probability distribution $\theta\Delta tp0$ and the steady state probability distribution (i.e., the lipotype frequency).

4.5.10. Data and Code Availability

All data are available in the main text and supplementary materials of this chapter, as well as those from the original, complete publication, are available online. Pipelines for cell segmentation with Cellpose are available at: https://doi.org/10.21228/M8698W and https://doi.org/10.5281/zenodo.6023316. Additional materials for digital phase contrast on primary dHFs can be found at: https://doi.org/10.5281/zenodo.5996882. Data, code, and Jupyter notebooks containing the implementation of CELLMA and to reproduce the results presented in this chapter are available at: https://doi.org/10.5281/zenodo.6245943. Further

information and requests for resources and reagents can be directed to the corresponding authors of Capolupo et al 2022.

4.6. Supplementary Materials

4.6.1. Supplementary Figures

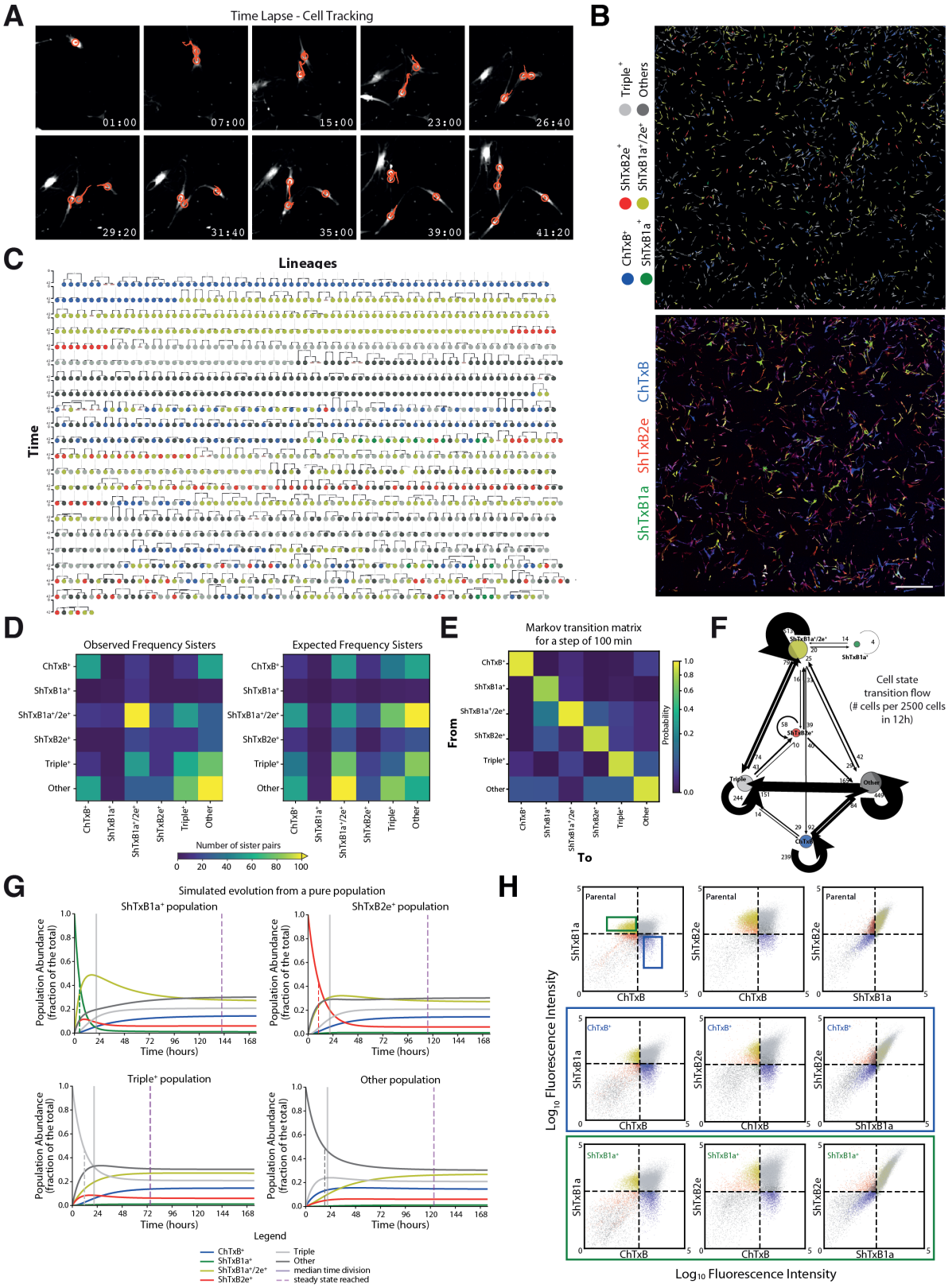


Figure S4.1. Cell tracking and lineage reconstruction from time-lapse recordings of dHFs. Adapted from Fig. S4 of Capolupo et al 2022. **(A)** Closeup of the time-lapse images illustrating an example of annotated tracking for a single dHF and its daughter cells over approximately 42 hours. Red circles indicate segmented cells and tailing lines indicate the tracked movement of cells. Time is

indicated in format HH:MM. **(B)** Images showing annotated lipotype call masks (upper panel) obtained by binarizing the information derived from the toxin staining (lower panel) at the endpoint of the movie. Lipotype masks are obtained by segmenting the last frame of the time-lapse experiment (also see Movie S1 from Capolupo et al 2022). Scale bar is 500 µm. **(C)** Reconstructed lineage hierarchy map of dHFs over 42 hours. Branches indicate intermediate cells (for which the state is latent) and circles indicate leave cells, colored by the observed lipotype class (green: ShTxB1a+, red: ShTxB2e+, yellow: ShTxB1a+/2e+, blue: ChTxB+, light grey: Triple+, dark grey: Other). **(D)** Frequency matrix of observed (left) and expected (right) sister cells lipotypes. Observed frequencies were obtained from C, expected frequency by randomizing the labels (**see Methods 4.3.9**). **(E)** CELLMA Markov transition matrix corresponding to a 100 min transition. **(F)** Graph of lipotype transition flow over 12 hours of the system at steady state. The number of cells transitioning between or within lipotypes is out of a total of 2,500 cells; edges corresponding to flows of less than 4 cells were pruned to simplify the interpretation. Cell numbers are located close to the corresponding arrow tail. **(G)** Simulated evolution of lipotype composition of pure ShTxB1a+, ShTxB2e+, Triple+, and Other populations over 7 days. An extension of Fig. 4.2G. **(H)** ChTxB+ and ShTxB1a+ cells were FACS-sorted and kept in culture for 10 days. Cells were stained with bacterial toxins and analyzed by cytofluorometry. Scatter plots of fluorescence intensity values are shown for each toxin. Populations are colored as in (A). Unstained cells were used as negative control to determine the gates. Blue and green boxes in CTRL cells indicate the FACS-sorted populations.

4.7. Appendix

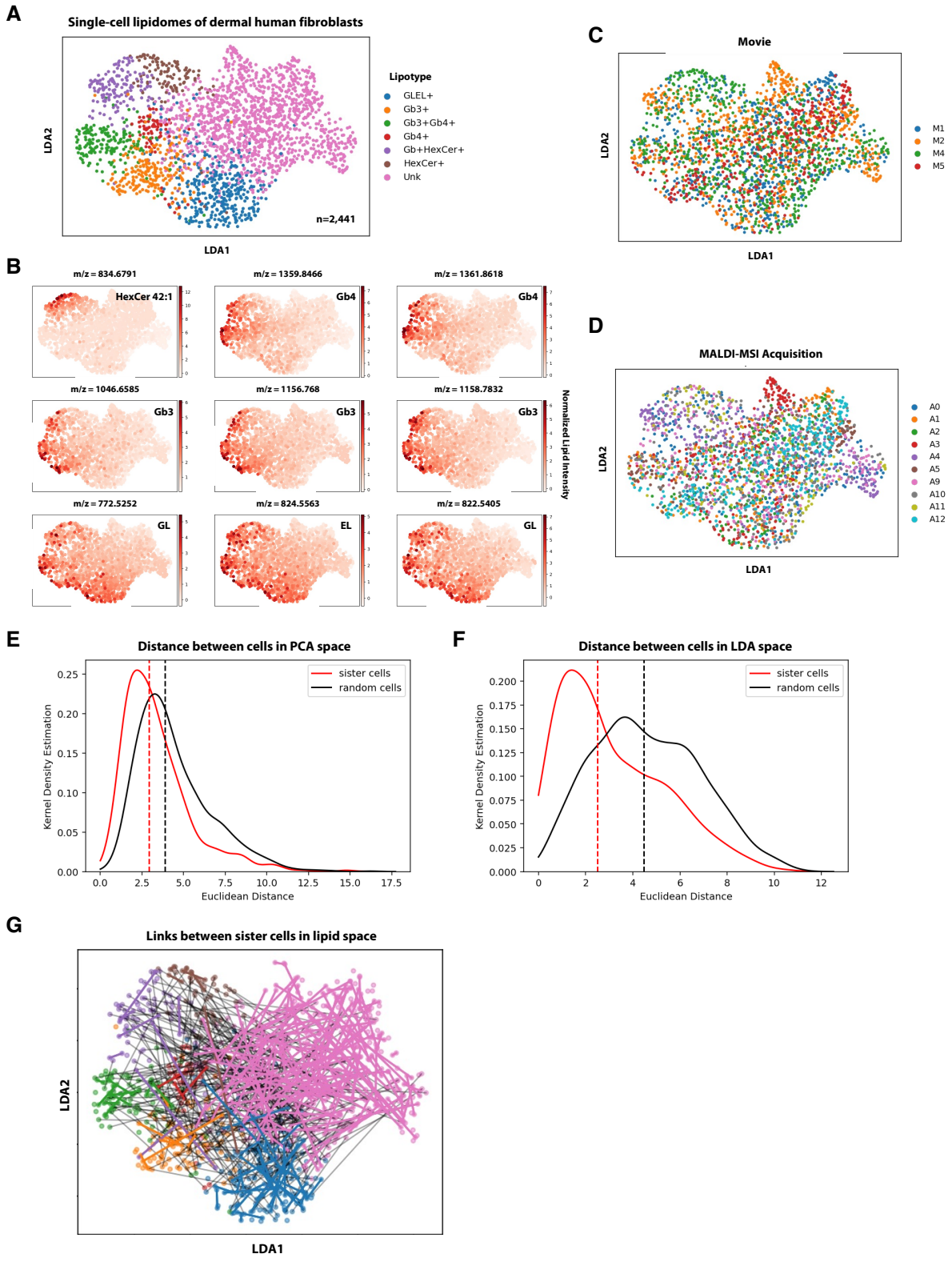


Figure A4.1. Mapping single-cell lipidomes of lineage reconstructed dHFs. (A) Linear discriminant analysis (LDA) of 2,441 single-cell dHF lipidomes obtained over four time-lapse movies and ten acquisitions with MALDI-MSI. Cells are colored by their lipotype determined using a feature space of 183 lipid molecules. **(B)** LDA plots showing the normalized intensities of nine variable lipids belonging various lipid classes (Gb3: trihexosylceramides, Gb4: globosides, HexCer: hexosylceramides, GL:

glycerolipids, EL: ether lipids). **(C-D)** LDA plot from (A) with cells colored by the time-lapse movie (C) or MALDI-MSI acquisition (D) from which they were acquired. **(E-F)** Kernel density plot of the Euclidean distance between a cell and its sister cell in PCA (E) or LDA (F) space compared to their distance with non-sister, random cells. **(G)** LDA visualization of single-cell lipidomes including links between related sister cells. Links between sister cells belonging to the same lipotype are colored according to that lipotype (see A for legend). Sister cells that do not belong to the same lipotype are instead linked with a black line.

5. Perspectives

In this thesis, I describe multifaceted efforts to decode the underlying properties of cell state transitions using quantitative analyses and modeling. My goal was to better understand how cells transition across cell states. I tackled this overarching idea with three specific biological research questions, each of which relates to the main dimensions of cell state transitions: their path, pace, and rules. In each chapter, I advance our scientific understanding of how cells transition across states and propose improved computational solutions to model those dynamics in a particular biological system.

In the first part, I tackle cell state transitions from a population perspective, asking which transitions unfold in a differentiation protocol where a homogeneous pool of progenitors is directed towards a particular cell fate. Specifically, I address this problem in a pragmatic setting: the characterization of dynamics during hESC-RPE differentiation for treatment of AMD (**Chapter 2**). I conclude that cells during the Reyes-Reurer protocol, rather than progressing along a direct linear path from stem cells to RPE, undergo a more complex dynamic that can be modeled as a divergence-convergence process, closely mirroring development. Motivated by this discovery, I further probe the plasticity of intermediate cell states by inducing alternate differentiation routes and, finally, confirm the identity of the final state reached by a detailed comparison with *in vivo* reference atlases and other published protocols.

In the second part, I consider the pace at which cell state transitions occur, asking how the rate of the cell cycle across different tissues and environmental contexts varies and whether it can be inferred with gene expression measurements from an ensemble of cells. To this end, I reformulate RNA velocity and gene manifold estimation into a unified framework, implementing a probabilistic model for velocity inference of periodic processes (**Chapter 3**). I describe changes in cell cycle speed in different *in vitro* and *in vivo* samples and propose a statistical tool for evaluating velocity significance. I also discover that cell cycle velocities can be approximated in real time and validated experimentally.

In the third part, I examine the role of cell state transitions to preserve a dynamic equilibrium of cell states in a biological system at homeostasis, asking whether the rules that govern transition probabilities among states can be defined using the partial information extracted from imaging-based lineage tracing. To explore this, I represent cell state transitions as Markov chains and create a model to estimate a transition probability matrix from reconstructed cell lineage information and endpoint lipid-state measurements (**Chapter 4**). I apply my method to investigate the fact that dHFs exist in different lipid compositional states.

My method uncovers the transition properties within this system at equilibrium, and I provide important evidence behind the concept of a lipotype: a cell state characterized and maintained because of its lipid composition.

These three chapters advance our collective understanding of stem cell therapies for retinal disease, cell cycle modulations, and lipid states in dermal fibroblasts. Naturally, there remain many open-ended questions that provide opportunities for further exploration. Here, I summarize the major findings of this thesis, dive into potential shortcomings, and contemplate avenues for future research. I also consider the interrelatedness of these works and describe how our understanding of cell state transitions continues to evolve.

5.1. Path: Molecular profiling of stem cell-derived retinal pigment epithelial cell differentiation established for clinical translation

5.1.1. Open questions regarding *in vitro* hESC-RPE differentiation

AMD is a major cause of vision loss and has significant promise to benefit from hESC-derived cell-replacement therapies. Among the different possible therapeutic approaches, one of the most feasible options involves the generation and transplantation of healthy RPE cells. In Chapter 2 of this thesis, we describe a 2D monolayer hESC-RPE differentiation protocol, leveraging scRNA-seq to assess intermediate cell states and provide a high-resolution perspective on a stem cell-based product intended for clinical trials. We transcriptomically characterize hESC-RPE differentiation at six time points and in three cell lines, demonstrating successful RPE lineage induction, selection, and maturation over 60 days. Initially, however, we observe a significant heterogeneity in the cell pool, with spatially patterned cell types resembling various sensory tissues adjacent to early optic tissues *in vivo* (**Fig 2.2**).

Whether the generation of these intermediate populations is necessary to obtain the final outcome of a pure RPE population is still unknown. While non-trivial experimentally, attempting to decouple the differentiation of these diverging populations would be an interesting research direction and help to understand the key factors fueling the process. For instance, the existence of these cells might support a diverse signaling environment, akin to the one present during embryonic development. Conversely, off-target populations may increase the time required to reach a mature RPE status. A diverse range of cell types also emerge in 3D embryoid body protocols and organoid systems, so the phenomenon is not restricted to more straightforward 2D differentiations (Azar et al., 2021). Nonetheless, this early-stage diversity, followed by a collapse onto a homogenous RPE cell population, supports a divergence-convergence model of *in vitro* differentiation (**Fig. 2.2 and 2.6**). It remains to be

seen whether such dynamics generalize to other protocols, as the tendency to profile time points besides those at the start and end is only recently becoming a more common practice.

Furthermore, our work highlights the role of intermediate cell stages in differentiation. We find that removing a specific progenitor cell population (NCAM1-High) at the protocol midpoint (D30) leads to a more rapid attainment of mature RPE (**Fig. 2.4**). Further efforts to characterize the population of cells that is low in levels of NCAM1 (i.e., CD140b-High cells; **Fig. S2.4**) may offer insights into the differentiation path that could assist with designing a shorter protocol, without compromising the quality or maturity of the final RPE product.

We also demonstrate that NCAM1-High populations can be re-directed at 30 days using another differentiation protocol to generate neuroepithelial cell types, including neuronal precursors (**Fig 2.5**). This finding could be promising because the second major cell type impacted by AMD are photoreceptor neurons. If it were possible to generate both RPE and photoreceptors from a single hESC precursor, or even better, to derive a functionally interacting co-culture of both cell types, it would be an enormous advancement towards actionable therapies for AMD. While most existing retinal differentiations produce only RPE or photoreceptors (Fortress et al., 2023; Rubner et al., 2022), the plasticity of intermediate cell states in our hESC-RPE protocol offers a chance to study the cues triggering commitment towards particular cell fates.

Despite reaching a high RPE purity after 60 days, a small portion of the final cell product still consists of retinal progenitor cells (**Fig 2.6**). Curiously, the same progenitor populations were detected in significantly longer protocols up to one year long (**Fig. S2.6F**) (Lidgerwood et al., 2021). This suggests a relationship between lingering progenitors and mature RPE cells. While it is typically assumed that RPE do not reverse their degree of differentiation, we cannot fully exclude that backwards transitions exist at low rates. These retinal progenitors do not seem to pose a safety risk, and they could benefit the mature RPE population by ensuring a renewed proliferative capacity in the event of environmental stressors or cell death. Perhaps the presence of a resident pool of progenitors is required to maintain a healthy culture in the first place. The study of these populations will likely yield valuable insights into the plasticity of final cultures.

5.1.2. Future challenges when preparing hESC-RPE for cell therapies

A critical aspect for the success of cell-based treatment therapies is transplantation into patients (Rizzolo et al., 2022). The three major concerns at this stage include the risk of de-differentiation into more pluripotent or unstable cell states, the ability of transplanted cells to functionally integrate with other host cell types, and the possibility of tissue rejection by the

host (S. Gupta et al., 2023). In the protocol we study, we do not observe any signs of de-differentiation or the presence of non-RPE cells among the endpoint cells grafted into the rabbit retina profiled by scRNA-seq and histological stainings (**Fig. 2.7**). Interestingly, the transplanted cells not only retained their RPE characteristics, but may have also further matured towards a more functional state, as indicated by the expression of visual cycle genes uniquely following transplantation. This suggests the cell state of transplanted RPE cells more closely matches that of resident normal cells in the adult.

The discovery that differentiation proceeds further after grafting opens the possibility that shorter differentiation protocols might be equally successful. While this next research direction is promising, in order to confidently determine whether RPE transplantations functionally support the host photoreceptor cells, transplantation experiments of RPE cell pools at different stages of differentiation need to be performed and assessed by scRNA-seq. This would ideally be accomplished in a rabbit model for retinal neurodegenerative disease. Finally, tissue rejection remains possible during therapy, but could be minimized by use of replacement tissues from patient-derived stem cells (Dehghan et al., 2022).

The viability of RPE after cryopreservation is important for clinical application of stem cell therapies. To evaluate whether long-term cryopreservation of these cells is possible, we are contributing to a follow up study in which we perform scRNA-seq on D60 cells that have replated for an additional three days, to enable viable cryopreservation for up to one year of storage. By comparing the transcriptome of cells at the protocol endpoint (not viable for cryopreservation, D60) to those thawed and re-cultured for three days (viable for cryopreservation, D60+3), we observed that preservable cells resembled more immature RPE (**Chapter 2.8, Fig. A2.1**). This was further confirmed by applying our ordinal classification scheme to newly collected data (**Fig. A2.1D**, **Fig. 2.7**). More experiments will need to be carried out to better understand the mechanism behind this de-differentiation of cell state and to evaluate overall cell viability. However, one hypothesis is that the small pool of retinal progenitors at D60 may promote RPE renewal after cryopreservation. The presence of increased EMT and proliferation signatures (**Fig. A2.1F-G**) supports this idea, since we also see these signatures after replating at D30 of the standard protocol, where NCAM1-High progenitor cells are present (**Fig. S2.5**).

Adapting stem cell therapies to the clinic will also pose other challenges. Generating large quantities of tissue to treat the large number of affected patients will require automation of differentiation protocols in clinical cell lines (i.e., KARO1 and E1C3). Moreover, other retinal diseases, such as those with a strong genetic basis, are not as advanced towards treatment with stem-cell based therapies; research using patient-derived tissue or organoid cultures is

a potential opening to address this (Liang et al., 2023). Nonetheless, significant efforts are ongoing to improve treatments for AMD and the first clinical trials have started (X. Chen et al., 2023). The efforts described here support a more nuanced characterization of cell state transitions during differentiation protocols with a strong practical intent to act as therapies.

5.2. Pace: Statistical inference with a manifold-constrained RNA velocity model uncovers cell cycle speed modulations

5.2.1. Improvements and limitations of existing RNA velocity methods

During the past five years, RNA velocity (La Manno et al., 2018) has emerged as an important technique in single-cell genomics for analyzing the rate of change during cell state transitions. As with the previous rise of pseudotime trajectory inference (Bendall et al., 2014; Trapnell et al., 2014), RNA velocity has inspired numerous computational biologists to develop extensions upon the original approach. Typically, these follow-up efforts focus on: (1) revising the differential kinetic equations to incorporate multiple omics modalities (Burdziak et al., 2023; Gorin et al., 2020; C. Li et al., 2022; S. Ma et al., 2020; Tedesco et al., 2022), (2) applying machine learning methods to obtain estimates that are better aligned with the biology (Bergen et al., 2020; Z. Chen et al., 2022; Cui et al., 2024; X. Qiu et al., 2022), or (3) adjusting the representation of latent time to address complex branching trajectories (M. Gao et al., 2022; Qiao & Huang, 2021).

These adjustments to RNA velocity estimation have been achieved due to progress in two related research domains. First, experimental techniques have been developed to jointly profile, in the same single cell, two or more omics modalities (i.e., transcriptome, chromatin accessibility, histone modifications, surface proteins). The joint detection of these modalities was in its early stages when the initial RNA velocity paper, with its implementation *velocyto*, was released. Second, advanced machine learning frameworks based on neural networks and generative modeling have been increasingly applied to estimate the kinetic parameters of the velocity equations and overcome the steady-state assumptions. Together, these general advancements have sought to improve interpretability and utility of the RNA velocity algorithm.

At the same time, there has been a broader recognition of the limits to existing velocity approaches and other caveats for analyzing single-cell data (Bergen et al., 2021; Gorin et al., 2022; Zheng et al., 2023). Bergen et al focuses on the problem of using the velocity rate equations to accurately model biological systems with more complex transcriptional dynamics, such as transcriptional boosting during erythroid maturation. In contrast, Zheng et al discusses the dependence of the RNA velocity workflow on k-nearest neighbors smoothing in order to

reduce noise. Finally, Gorin et al provides a comprehensive overview of the entire RNA velocity workflow, including preprocessing, model estimation, and embedding; they conclude that the route to improving future velocity methods is design more tractable inference procedures, such as with Bayesian inference, rather than extend the capabilities of existing models with additional heuristics.

5.2.2. Intronic transcript detection levels with short-read sequencing

Due to low molecule detection levels, gene expression count data is highly sparse, and poor detection of intronic reads makes velocity estimation challenging. Certain biases in intronic read detection are also present when using short-read Chromium 10X sequencing, and these may affect velocity estimation. Additionally, a large amount of internal priming in repetitive A-rich nucleotide regions occurs, and this has been demonstrated to amplify intronic reads in a highly biased manner (10XGenomics, 2020).

Recent gene expression quantification algorithms have been developed to improve memory-efficiency, reduce false-positive molecule detection, and account for reads spanning multiple genes (D. He et al., 2022; Kaminow et al., 2021; Melsted et al., 2021). However, these computational methods are limited because most scRNA-seq data is generated without sequencing the entire transcript body, preventing reliable detection of intronic regions and isoforms. These approaches also do not address the technical biases of 10X protocols

Improved quantitative UMI technologies with full-length transcript sequencing, such as Smart-seq3xpress and FLASH-seq, better discriminate among alternatively-spliced variants of a gene, but these technologies are low-throughput and expensive (Hagemann-Jensen et al., 2022; Hahaut et al., 2022). Long read sequencing technologies, once adapted for single cells and perhaps paired with a modified Chromium 10X protocol, could yield improvements to unspliced and spliced RNA molecule detection (Marx, 2023; Tian et al., 2021).

5.2.3. Data preprocessing and RNA velocity benchmarking metrics

Aspects of data preprocessing offer computational challenges that ought to be more widely considered when creating new velocity tools. For example, the use of nearest-neighborhood imputation and other data smoothing techniques to overcome sparsity may unintentionally cause signal to bleed from some genes to others, distorting the low dimension embedding upon which RNA velocity vector fields are projected (Gorin et al., 2022). Likewise, somewhat arbitrary decisions about dimensionality reduction parameters and non-geometric embedding strategies (i.e., tSNE, UMAP) risk distorting the widely used two-dimensional vector field representation format for velocity (Chari & Pachter, 2023).

Furthermore, there are few standardized metrics to evaluate if a velocity is trustworthy. Some initial benchmarking scores for velocity have been formulated, including the consistency score, cross-boundary direction correctness, in-cluster coherence, and velocity coherence (Bergen et al., 2020; M. Gao et al., 2022; Gayoso et al., 2023; Qiao & Huang, 2021). However, these metrics have limitations that reduce the scenarios in which they can be applied to data.

Velocity consistency and cross-boundary direction correctness evaluate how likely a particular cell, given its velocity, is to reach the target cell type (Bergen et al., 2020; Qiao & Huang, 2021). This is achieved by comparing the vector field of the cell to its nearest-neighbors (velocity consistency) or to nearby cells in the low-dimensional embedding space (cross-boundary direction correctness). While these two approaches will indicate whether a velocity vector aligns to a specific differentiation trajectory, they require a ground truth direction. Therefore, neither metric can be applied to samples with an unknown biology, including when comparing a diseased or perturbed system to a control. In these circumstances, one would be unable to conclude whether a low metric score is due to a biological change in trajectory or just poor-quality data.

In-cluster coherence (Qiao & Huang, 2021) does not require a ground truth direction and is computed as a cosine similarity between cell velocities in the same cluster; a higher similarity indicates a less noisy velocity. However, this metric assumes appropriate cell clustering and scrutinizes the low-dimensional velocity projection, not the high-dimensional space in which velocity itself is estimated.

Finally, velocity coherence (Gayoso et al., 2023) is a gene-wise score that examines how well a gene's velocity agrees with that of the cell-wise estimate, which contains information across all genes as well as from similar cells used for nearest-neighbor smoothing. This metric will indicate if a particular gene's velocity does not align with that of the average across all genes; however, this is not necessarily due to poor velocity estimation and could be biologically meaningful. By extrapolating a velocity estimate across all genes, the obtained vector field represents a multitude of biological signatures (i.e., cell cycle, differentiation, development, stress response), each defined by a unique set of genes. For example, in a population of differentiating stem cells undergoing a transformation towards a mature cell type (**Chapter 2**), one might expect both cell cycle and differentiation programs to co-occur. A velocity across all genes effectively combines these two signatures into a single vector field, even though estimates obtained solely with cell cycle or differentiation genes would yield different vector directionality. Thus, when applying the velocity coherence metric, the particular gene being evaluated could indicate a low coherence simply because it is affiliated to a minor (i.e., cell cycle) rather than major (i.e., development) axis of variation.

In terms of rigorous statistical testing, *VeloCycle* is the first method that allows direct statistical comparison and evaluation of velocity differences between samples, using the uncertainty measurement obtained from Bayesian inference procedures (**Fig. 3.6**). Other exciting probabilistic velocity methods have been recently proposed (Aivazidis et al., 2023; Gayoso et al., 2023; Gu et al., 2022; Qin et al., 2022), but these approaches mainly apply uncertainty quantification to assess the trustworthiness of a velocity estimate rather than to directly compare samples or conditions. Evaluating differential velocity among biological samples is possible in *VeloCycle* due to joint estimation of the manifold and velocity.

5.2.4. Applicability of RNA velocity to single-nuclei data

Single-nucleus RNA sequencing data (snRNA-seq) is obtained by isolating the RNA content present within a cell's nucleus; this procedure discards any RNA molecules in the cytoplasm. Libraries for snRNA-seq are easier to obtain than for scRNA-seq when working with freshly frozen tissues and using protocols that jointly profile two omics modalities, such as 10X Multiome. The construction of transcriptomic atlases using human tissues from adult donors or embryos, for which the time between sample collection and preparation varies, single-nuclei protocols are often the only available option (Kim et al., 2023).

RNA velocity has been successfully applied to snRNA-seq datasets (Kang et al., 2023; Marsh & Blelloch, 2020; Wolfien et al., 2020); however, since a significantly larger fraction of detected UMIs correspond to unspliced rather than spliced RNA (50-70% unspliced reads in single nuclei compared to about 30% unspliced reads in single cells), it is unclear whether the assumptions about the velocity kinetic parameters hold (10XGenomics, 2020; Alvarez et al., 2020). Specifically, it is the much faster nuclear export rate, rather than the degradation rate, that is likely modeled during velocity estimation on single nuclei. In the future, a systematic comparison of single-cell and single-nucleus RNA velocities obtained on the same cell population would help elucidate any differences in performance. With *VeloCycle*, one could estimate the cell cycle speed on single-nucleus and single-cell data from the same cell culture, and evaluating whether different estimates of the velocity function and kinetic parameters are obtained. With appropriate scaling of the priors, one could perhaps adjust the "degradation rate" to be on a scale of the faster "nuclear export rate" detected in single-nuclei data.

5.2.5. Improvements of a manifold-consistent RNA velocity

While it may not be possible to resolve all of the challenges with RNA velocity estimation in a single framework, remaining cognizant of the current limitations of velocity approaches is important for guiding the future direction of method development, and it will help

the community to best identify scenarios in which velocity can nonetheless be informative for biological discovery.

VeloCycle reformulates RNA velocity such that the vector field is explicitly defined on the coordinates of the gene expression manifold. The major benefit of this approach is that it ensures geometric consistency; the velocity always points in a direction that, when followed, leads to a tangent space that is spanned by the data. In other words, there is no risk of velocity projecting a cell's future state to a position completely outside the gene expression manifold. The velocity function in *VeloCycle* is also a cell- and gene- independent entity, addressing the problem of gene-wise level velocities, which measure the rate of change at different time scales (**Fig. 3.1**). We implement this reformulation as a probabilistic framework intended for the one-dimensional cell cycle manifold, which enables statistical testing of RNA velocity estimates by examining the overlap between credibility intervals (**Fig. 3.7**). For the first time, it is also possible to compare computationally inferred velocity estimates to those measured by experimental techniques, such as time-lapse microscopy (**Fig. 3.6**). *VeloCycle* is a modular tool, operates on raw counts, and does not involve projection of a vector field onto non-geometric embeddings. These distinct features will hopefully motivate future velocity model development towards more biophysically tractable frameworks.

5.2.6. Automation of gene selection procedures with *VeloCycle*

Despite the achievements of *VeloCycle*, there are possible improvements that could enhance model performance and interpretability (**Chapter 3.4**). *VeloCycle* does not conduct unbiased gene selection beyond initial quality control and uses well-characterized marker genes from literature sources (Ontology Consortium et al., 2023; Riba et al., 2022; Satija et al., 2015). Since only a subset of all genes are actually expressed in a periodic manner along the cell cycle manifold, the lack of an automated gene selection framework requires a choice between two options. In the first, a large and noisy set of genes is used in training, but the model may fail to converge. In the second, a small but literature-based set of genes is used in training, which might fit a more accurate velocity; however, the model may overfit and could be sensitive to variation between biologically similar datasets with different amounts of noise. Initial efforts in *VeloCycle* to optimize gene selection and remove noisy genes in an unbiased manner includes removing genes with a low cross-correlation or a negative phase delay between peak unspliced-spliced levels (**Chapter 3.5**).

In the future, one approach to improve gene choice would be to incorporate feature selection itself into the *manifold-learning* or *velocity-learning* procedures. As a part of *manifold-learning*, one could define a Bernoulli random variable that indicates, for each gene, a certain

probability that it is expressed periodically along the cell cycle. A gene would be considered non-periodic if the loss obtained with a constant-value function to fit spliced expression is lower than the loss with a function defined by multiple harmonics. Models formulated in this way, sometimes referred to as Latent Bernoulli Allocation, would facilitate an independent identification of genes fluctuating along the periodic process. One downside to this approach is that selecting strong priors for the Bernoulli random variable biases the model towards a particular proportion of all genes being periodic. On the contrary, if the prior is too weak, even noise might be biased to be selected, distilling the feature selection problem to the tuning of metaparameters that are difficult to set in a balanced fashion. Alternatively, gene selection could be performed in *velocity-learning*. One could, for example, consider a null velocity model in which the gene-wise fit for unspliced expression is computed using a constant of zero velocity. For genes where the unspliced estimate using the null model yields a lower error between the expected and observed fits, there is limited velocity information present and the gene should be excluded from angular speed inference. Another strategy would be to select genes based on the uncertainty correlation between the degradation rate and angular speed. As observed by MCMC and the SVI+LRMN reformulation (**Fig. 3.5**), there is an underlying correlation structure between the velocity and the kinetic parameters for some, but not all, genes. This alludes to a technique for discriminating between “velocity-informative” (correlated uncertainty) and “velocity-uninformative” (uncorrelated uncertainty) genes.

5.2.7. Continuous formulation of the kinetic parameters with *VeloCycle*

Another enhancement to *VeloCycle* would be to reformulate the kinetic parameters (splicing rate and degradation rate) such that they are non-constant functions evolving along the cell cycle manifold. This representation would more closely reflect the underlying biology, in which splicing and degradation rates are expected to vary at each phase of cell division (Battich et al., 2020; Mizukoshi et al., 2023). One could speculate that splicing and degradation rates should be parameterized similarly to spliced gene expression using a truncated Fourier series with one or more harmonics. However, it is unclear whether a periodic representation would be sufficient to model multiple fluctuations in kinetic rates during a single cell cycle. Maintaining a well-conditioned model while achieving non-constant representations of the kinetic parameters is also non-trivial because adding two additional gene-specific sets of Fourier series components would increase model complexity. Without proper priors, perhaps obtained biologically, the model could fail to converge on a reasonable estimate. To achieve this in *VeloCycle*, a careful study of the priors for splicing and degradation rates across cell types and conditions should be performed first.

5.2.8. Evaluating changes to chromatin accessibility and velocity with *VeloCycle*

The incorporation of multiple modalities into RNA velocity models using extensions of the original system of ordinary differential equations could enable the inference of the "acceleration" of a cell's velocity along a particular trajectory (Lederer & La Manno, 2020). Such concepts have been previously implemented using joint profiling of RNA and cell surface proteins (Gorin et al., 2020) as well as of RNA and chromatin accessibility (C. Li et al., 2022; S. Ma et al., 2020; Tedesco et al., 2022). These approaches suffer the same limitations as other velocity models, including the use of gene-wise velocity estimates and no requirement that velocity remains tangent to the manifold.

Therefore, it would be interesting to extend *VeloCycle* to incorporate the use of multi-omics data. The cell cycle is regulated by transcription factors and other changes to chromatin structure, particularly in S phase during DNA replication (Y. Ma et al., 2015). The *manifold-learning* procedure alone could be insightful to identify changes to chromatin accessibility that unfold during the cell cycle progression in varying cellular contexts. Moreover, since multi-omic data jointly measuring gene expression and chromatin accessibility is always achieved at the single-nuclei level, this could be an interesting setting to assess the performance of single-nuclei in velocity estimation.

Changes in chromatin accessibility have also been shown as important in the decision-making process for cell cycle exit (Y. Ma et al., 2019). Another future avenue of work would be to develop an approach that could better discern between G1 and non-proliferative G0 stages, which is challenging using transcriptomic level measurements alone (Oki et al., 2014; Theilgaard-Mönch et al., 2022).

5.2.9. Manifold-constrained velocity for non-periodic biological systems

VeloCycle is a specially tailored implementation of manifold-constrained velocity estimation for systems with periodic gene expression manifolds. However, the proposed mathematical framework is not restricted to the cell cycle or one-dimensional manifolds, and it could be applied to estimate velocity along linear or branching trajectories. Although closely intertwined with certain neurodevelopment programs, including for radial glial progenitors (Alieh et al., 2023), the cell cycle is admittedly not usually the primary axis of variation when characterizing a tissue or cell population with single cell transcriptomics. Hence, it would be valuable to extend or rework the model from *VeloCycle* to be applicable to temporal progression in gene expression space that corresponds to either cell differentiation or development.

Significant efforts will be needed to appropriately parametrize higher dimensional manifolds: unlike one-dimensional periodic manifolds, which can be conveniently modeled using a truncated set of Fourier series components, the parameterization choice for more complex manifolds is non-trivial. For example, parameterizing a trajectory branching point might require incorporating a mixture of two one-dimensional manifolds. To overcome this, more intricate combinations of gene expression trajectories could be broken down into individual, but one-dimensional, trajectories. One of these trajectories would represent a single axis of variation in the data, such as progression from progenitor X to mature cell type Y, along which gene expression could be modeled using an exponential function or linear spline. However, this implies modeling both genes and cells as disjoint sets, forcing the challenging identification of which genes vary along a particular manifold. Thus, comparisons of the same axis of variation across multiple conditions would be limited without some sort of common trajectory definition. In the long term, RNA velocity frameworks that consider multiple manifolds with varying topologies, traversed by cells in different gene subspaces and giving rise to multiple velocities (i.e., cell cycle velocity, cell differentiation velocity), would offer a broader utility for manifold-constrained, gene-independent RNA velocity estimation.

5.3. Rules: Charting recurring cell lipid-state transitions with lineage leaf-state Markov analysis

5.3.1. Challenges with the Markov formulation of CELLMA

Compared to linear processes in biology, recurring cell state dynamics are less frequently studied with single-cell technologies. One reason is that modeling these systems, in which a cell may visit a particular state multiple times and in an unordered manner, difficult with static snapshots and only possible with samples at a steady-state. Methods to extrapolate future cell states, such as RNA velocity, only indicate the rate at which cells are changing and do not offer insight into the underlying rules that control those transitions.

RNA velocity extracts information from RNA metabolism, which occurs on a time scale similar to that of cell state transitions. Similarly, CELLMA uses cell proliferation, which also happens on a time scale similar to that of cell state conversions, to understand the rules of state transitions modeled as discrete entities. CELLMA elucidates the relationships among cell states by estimating a transition matrix containing the probability that one state will transition into another during a particular time frame. These states are defined by synthesized lipids, rather than the transcriptome, and transitions are inferred from time-lapse microscopy

imaging and end-state toxin stainings. We demonstrate that the transitions between lipid-states, which we refer to as lipotypes, can be modeled in dHFs (**Chapter 4**).

We chose to represent dHF cell state transitions as a memoryless Markov model because the biological system is at steady-state. In other words, there is a fixed proportion of cells with each of the different lipotypes; transitions do occur, but always in a way that maintains the overall proportions in the cell culture. One challenge with this implementation is that there are covariances between entries of the transition matrix. The is related to the constraint that for the matrix to be valid Markov matrix, the sum of each rows must equal 1, with all entries between 0 and 1. Likewise, conditioned on a given steady state, a change the probabilities in one row will impact the probability values in other rows. Consequently, it is possible for probabilities to be distributed differently in the matrix without impacting the global lipotype proportions. For example, in a three-state system, the outflow of cells from “state 1” could be 2% at a given time step, but the distribution of that outflow to “state 2” and “state 3” can be achieved using different splits (i.e., 1%/1%, 0.5%/1.5%, or 0.1%/1.9%). Further efforts are needed to assess CELLMA performance on simulated data with a ground truth transition matrix.

5.3.2. Reconstructing cell lineages with time-lapse microscopy

Lineage tracing in single cells is an active area of research (C. Chen et al., 2022; Wagner & Klein, 2020). As previously discussed (**Chapter 1**), several experimental techniques have been created to molecularly record cell lineages and clonal history using diverse approaches, including viral barcoding (Kong et al., 2020; Weinreb et al., 2020) and the CRISPR/Cas9 system (Alemany et al., 2018; Chow et al., 2021; McKenna et al., 2016; F. Schmidt et al., 2018). While promising, these tools are challenging to adopt and not necessarily suitable to the biological system being studied. Tracking lineages by time-lapse microscopy remains amenable to many cell types, and it allows direct reconstruction of lineages as opposed to an indirect inference. CELLMA demonstrates that cleverly designed models can use time-lapse images to uncover novel insights with fewer technical obstacles.

Two challenges when working with time-lapse imaging data are cell segmentation and tracking. Cell segmentation was a notoriously difficult problem in biological image processing; however, the recent deep learning methods such as Cellpose and stardist perform well and have significantly improved segmentation quality, particularly in crowded cell culture settings (Pachitariu & Stringer, 2022; Weigert et al., 2020).

On the contrary, tracking cells in a time-lapse movie and linking them to their division events remains a difficult and unsolved problem. Tools such as TrackMate and ultrack have

features to account for cell splitting (Bragantini et al., 2023; Ershov et al., 2022), but crowded culture conditions and segmentation gaps can still accumulate and impact tracking. Cell markers, including DAPI nuclei staining and FUCCI signals that mark phases of the cell cycle, are not visible for a short period of time when a cell is dividing, which may contribute to gaps in cell lineages when applying the methods above. Ultimately, advances in the domains of image segmentation and tracking are especially crucial for single-cell lipidomics methods, which are image-based and, unlike transcriptomics, have no sequencing-based alternative.

5.3.3. Incorporation of lipotypes defined by MALDI-MSI into CELLMA

CELLMA models lipotype transitions defined by toxin stainings (ShTxB1a, ShTxB2e, and ChTxB), which effectively discriminate some lipid classes from each other (Gb3, Gb4, and GM1). However, there are other families of lipids, including hexosylceramides (Hex-Cers) that these toxin stainings cannot distinguish. Instead, applying CELLMA to model lipotypes defined by a combination of individual lipid measurements obtained using MALDI-MSI, rather than from toxin stainings directly, would offer greater resolution of full lipid heterogeneity in dHFs.

We have already collected thousands of single-cell lipidomes from multiple time-lapse movies using MALDI-MSI at the end-state readout. Preliminary analyses confirm that some of the lipotypes detected correspond directly with toxin-stained Gb3 and Gb4 classes (**Fig. A4.1A-D**). We also observe that sister cells tend to be more similar to each other in their lipid compositions and are spatially in closer proximity in PCA space compared to unrelated cells (**Fig. A4.1E-G**). Furthermore, we detect the presence of a Hex-Cers lipotype that was not observable with toxin staining (**Fig. A4.1A**). A next step would be to assess the performance of the original CELLMA model using these newly-defined lipotype classes.

Although MALDI-MSI method is promising for the study of lipid heterogeneity in single cells, there remain numerous technical challenges. Peak detection in the pixels of MALDI images can be disturbed by noise, and two lipids located in a similar m/z range may have overlapping detection windows. Moreover, measurements are highly sparse, sometimes even more than transcriptomics data. MALDI-MSI also measures molecule intensity, rather than a direct quantitative count of molecules. Therefore, assumptions made for scRNA-seq data during normalization and feature selection may not hold for these data and should be carefully re-evaluated. Likewise, batch effect correction between different regions of an image or across multiple acquired images poses an unsolved challenge. Currently applied methods have been repurposed from other research domains, such as Combat correction for microarrays (Balluff et al., 2021; Johnson et al., 2007), and were not designed with the specific data format in mind.

5.3.4. Modeling lipotype transitions in a co-culture setting

In the human body, dermal fibroblasts can exist in either a papillary or reticular state depending on their proximity to keratinocyte cells of the epidermis. Both states are characterized by distinct morphological features and gene expression patterns, but knowledge of the role lipids play in defining these states remains limited (B. Russo et al., 2020; Werner et al., 2007). Therefore, it would be interesting to track fibroblasts over time in a co-cultured setting with keratinocytes and model lipotype changes with MALDI-MSI. However, application of CELLMA to this co-cultured system would be challenging because the steady-state assumption of lipotype transition probabilities is likely violated. To overcome this, one could extend CELLMA to infer two transition steady-state matrices: one that defines the behavior for fibroblasts located in close (papillary steady-state) or far (reticular steady-state) proximity to keratinocytes. A new variable that incorporates cell distances and density would also be needed to determine when to switch between transition matrices during estimation. However, this requires either (1) an arbitrary choice of the relevant switching point between transition matrices (perhaps based on a cell distance threshold), or (2) the introduction of this parameter as a latent variable, which might significantly complicate inference.

5.3.5. The heritability of lipotype state transitions

Our work also reveals that stable lipotype states are conserved through cell divisions. Sister cells and even cousin cells were found to be more likely to have a more similar lipotype compared to cells from another clade, alluding to a heritability of lipid configurations (**Fig. 4.2**). This warrants future studies on the heritability of lipotypes across generations, as well as whether they are more stable states compared to those that are transcriptomically-defined. To this end, CELLMA could be used to model cell state transitions defined by another endpoint readout, such as single-cell spatial transcriptomics (Vandereyken et al., 2023).

5.4. Concluding remarks

The three works presented here harness technologies with single-cell resolution to investigate and model cell state transitions in unique research domains. Despite the different applications of these tools in each chapter, there are some valuable shared themes. One major unifying theme across these works is their novel application of single-cell methods to decode how cells change over time in response to intrinsic and extrinsic factors. Single-cell technologies are a resourceful means for researchers to explore dynamic systems (**Chapter 1.4**), and they will continue to be at the forefront of biological discovery for some time.

5.4.1. Recurring cell states in non-linear biological processes

Interestingly, another shared theme is the prominence of recurring cell states. Typically, most research on cell states is focused on one-way axes of biological change, with a clear beginning (progenitor) and end (mature) state. In these systems, cells are assumed not to transit though a particular cell state more than once: after a progenitor cell reaches a specific intermediate state, it is not expected to go backwards. The research discussed here, however, illustrates an unrealized prominence of non-canonical cell state transition routes in the single-cell genomics field.

For instance, stem cell differentiation protocols are designed with the clear intention to induce a specific cell fate. Thus, it is expected that hESCs evolve through intermediate states that steadily alter the transcriptomic profile of the cells closer towards that of the target cell type. Surprisingly, we observe that hESCs follow a non-linear trajectory, oscillating between various cell states and even undergoing de-differentiation in response to experimental steps such as replating.

Conversely, the cell cycle is quite obviously recurring. However, in single-cell datasets, it is often overlooked and even dismissed as a nuisance, interfering with a more insightful biological signature. We highlight that variations in cell cycle speed between samples with different environmental or genetic contexts can actually offer meaningful biological rationale that may complement time-series analysis of a primary linear trajectory such as differentiation. *VeloCycle* should remind the community that it is valuable to extract all axes of variation present in single-cell datasets, rather than regressing out a secondary process that is assumed to behave uniformly and uninformatively.

Finally, we demonstrate that recurring cell states are not always transcriptomically defined and can even occur in mature cell types such as dHFs. These steady-state lipotypes can be described using Markov models, which may seem counterintuitive to the standard view of cell state transitions that assumes a cell must remember the state it came from in order to advance to a future state. Here, we show that cell transitions can indeed be modeled without direct knowledge of previous states when those transitions are equally likely at any given time.

5.4.2. Modeling in single cell biology

Another shared theme is the importance of computational models in single-cell genomics. Single-cell data is high-dimensional and it can be difficult to extract biological meaning; thus, models are needed that effectively constrain the data to a well-posed structure, but without overparameterization. Well-reasoned assumptions can yield models with sufficient

inductive bias that offer new insights into familiar problems. For example, we develop an ordinal classifier to map hESC-RPE cells to a developmental maturity status. The machine learning field offers many types of categorical classification schemes, many of which could be performing well at this task. However, by incorporating the knowledge that differentiating cell classes are chronological, we propose an elegant and novel solution with good performance. This is also true by repurposing Markov chains, rarely used to model single-cell data, for defining lipotype transition matrices. Finally, with our reformulated framework for cell cycle velocity, we illustrate the importance of coupling biophysically reasoned models and uncertainty estimation to advance an RNA velocity field that has grown accustomed to relying on the original, but flawed, formulation.

5.4.3. Versatility and lasting potential of single-cell omics approaches

Taken together, this thesis demonstrates the immense versatility of single-cell omics data to facilitate diverse scientific accomplishments. The single-cell community often discusses the challenges of working with single-cell data, and the trend of atlasing is often attacked being shallow; however, it is worth reflecting on the possibility-opening value of these techniques, which enable us to ask scientific questions otherwise out of reach. Single-cell data is highly adaptable and, unlike many other types of collected biological data, can be repurposed time and time again to drive new discoveries. This flexibility of single-cell data to answer many biological questions is likely because it is, for now, the best measurements we have to access the dynamics of cell states. For this same reason, single-cell technologies will continue to be an essential method in the study of biological systems for many years to come.

References

10X Genomics. (2020). *Interpreting Intronic and Antisense Reads in 10x Genomics Single Cell Gene Expression Data*.

Adler, M., Mayo, A., Zhou, X., Franklin, R. A., Meizlish, M. L., Medzhitov, R., Kallenberger, S. M., & Alon, U. (2020). Principles of cell circuits for tissue repair and fibrosis. *IScience*, *23*(2), 100841. https://doi.org/10.1016/j.isci.2020.100841

Ahn, J., Heo, S., Lee, J., & Bang, D. (2021). Introduction to single-cell DNA methylation profiling methods. *Biomolecules*, *11*(7), 1013. https://doi.org/10.3390/biom11071013

Aibar, S., González-Blas, C. B., Moerman, T., Huynh-Thu, V. A., Imrichova, H., Hulselmans, G., Rambow, F., Marine, J.-C., Geurts, P., Aerts, J., van den Oord, J., Atak, Z. K., Wouters, J., & Aerts, S. (2017). SCENIC: single-cell regulatory network inference and clustering. *Nature Methods*, *14*(11), 1083–1086. https://doi.org/10.1038/nmeth.4463

Aissa, A. F., Islam, A. B. M. M. K., Ariss, M. M., Go, C. C., Rader, A. E., Conrardy, R. D., Gajda, A. M., Rubio-Perez, C., Valyi-Nagy, K., Pasquinelli, M., Feldman, L. E., Green, S. J., Lopez-Bigas, N., Frolov, M. V., & Benevolenskaya, E. V. (2021). Single-cell transcriptional changes associated with drug tolerance and response to combination therapies in cancer. *Nature Communications*, *12*(1), 1628. https://doi.org/10.1038/s41467-021-21884-z

Aivazidis, A., Memi, F., Kleshchevnikov, V., Clarke, B., Stegle, O., & Bayraktar, O. A. (2023). Model-based inference of RNA velocity modules improves cell fate prediction. In *bioRxiv* (p. 2023.08.03.551650). https://doi.org/10.1101/2023.08.03.551650

Albergante, L., Mirkes, E. M., Chen, H., Martin, A., Faure, L., Barillot, E., Pinello, L., Gorban, A. N., & Zinovyev, A. (2018). Robust And Scalable Learning Of Complex Dataset Topologies Via Elpigraph. In *arXiv [cs.LG]*. arXiv. http://arxiv.org/abs/1804.07580

Alemany, A., Florescu, M., Baron, C. S., Peterson-Maduro, J., & van Oudenaarden, A. (2018). Whole-organism clone tracing using single-cell sequencing. *Nature*, *556*(7699), 108–112. https://doi.org/10.1038/nature25969

Alieh, L. H. A., Herrera, A., & La Manno, G. (2023). Heterogeneity and developmental dynamics of mammalian neocortical progenitors. *Current Opinion in Systems Biology*, *32–33*, 100444. https://doi.org/10.1016/j.coisb.2023.100444

Alon, U., Barkai, N., Notterman, D. A., Gish, K., Ybarra, S., Mack, D., & Levine, A. J. (1999). Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proceedings of the National Academy of Sciences of the United States of America*, *96*(12), 6745–6750. https://doi.org/10.1073/pnas.96.12.6745

Alvarez, M., Rahmani, E., Jew, B., Garske, K. M., Miao, Z., Benhammou, J. N., Ye, C. J., Pisegna, J. R., Pietiläinen, K. H., Halperin, E., & Pajukanta, P. (2020). Enhancing droplet-based single-nucleus RNA-seq resolution using the semi-supervised machine learning classifier DIEM. *Scientific Reports*, *10*(1), 11019. https://doi.org/10.1038/s41598-020-67513-5

Ambati, J., Ambati, B. K., Yoo, S. H., Ianchulev, S., & Adamis, A. P. (2003). Age-related macular degeneration: etiology, pathogenesis, and therapeutic strategies. *Survey of Ophthalmology*, *48*(3), 257–293. https://doi.org/10.1016/s0039-6257(03)00030-4

Arai, Y., Pulvers, J. N., Haffner, C., Schilling, B., Nüsslein, I., Calegari, F., & Huttner, W. B. (2011). Neural stem and progenitor cells shorten S-phase on commitment to neuron production. *Nature Communications*, *2*, 154. https://doi.org/10.1038/ncomms1155

Auerbach, B. J., FitzGerald, G. A., & Li, M. (2022). Tempo: an unsupervised Bayesian algorithm for circadian phase inference in single-cell transcriptomics. *Nature Communications*, *13*(1), 6580. https://doi.org/10.1038/s41467-022-34185-w

Azar, J., Bahmad, H. F., Daher, D., Moubarak, M. M., Hadadeh, O., Monzer, A., Al Bitar, S., Jamal, M., Al-Sayegh, M., & Abou-Kheir, W. (2021). The use of stem cell-derived organoids in disease modeling: An update. *International Journal of Molecular Sciences*, *22*(14), 7667. https://doi.org/10.3390/ijms22147667

Balluff, B., Hopf, C., Porta Siegel, T., Grabsch, H. I., & Heeren, R. M. A. (2021). Batch effects in MALDI mass spectrometry imaging. *Journal of the American Society for Mass Spectrometry*, *32*(3), 628–635. https://doi.org/10.1021/jasms.0c00393

Baran, Y., Bercovich, A., Sebe-Pedros, A., Lubling, Y., Giladi, A., Chomsky, E., Meir, Z., Hoichman, M., Lifshitz, A., & Tanay, A. (2019). MetaCell: analysis of single-cell RNA-seq data using K-nn graph partitions. *Genome Biology*, *20*(1), 206. https://doi.org/10.1186/s13059-019-1812-2

Bartosovic, M., & Castelo-Branco, G. (2023). Multimodal chromatin profiling using nanobody-based single-cell CUT&Tag. *Nature Biotechnology*, *41*(6), 794–805. https://doi.org/10.1038/s41587-022-01535-4

Bartosovic, M., Kabbe, M., & Castelo-Branco, G. (2021). Single-cell CUT&Tag profiles histone modifications and transcription factors in complex tissues. *Nature Biotechnology*, *39*(7), 825–835. https://doi.org/10.1038/s41587-021-00869-9

Bartuma, H., Petrus-Reurer, S., Aronsson, M., Westman, S., André, H., & Kvanta, A. (2015). In Vivo Imaging of Subretinal Bleb-Induced Outer Retinal Degeneration in the Rabbit. *Investigative Ophthalmology & Visual Science*, *56*(4), 2423–2430. https://doi.org/10.1167/iovs.14-16208

Bastidas-Ponce, A., Tritschler, S., Dony, L., Scheibner, K., Tarquis-Medina, M., Salinno, C., Schirge, S., Burtscher, I., Böttcher, A., Theis, F. J., Lickert, H., & Bakhti, M. (2019). Comprehensive single cell mRNA profiling reveals a detailed roadmap for pancreatic endocrinogenesis. *Development* , *146*(12). https://doi.org/10.1242/dev.173849

Battich, N., Beumer, J., de Barbanson, B., Krenning, L., Baron, C. S., Tanenbaum, M. E., Clevers, H., & van Oudenaarden, A. (2020). Sequencing metabolically labeled transcripts in single cells reveals mRNA turnover strategies. *Science*, *367*(6482), 1151–1156. https://doi.org/10.1126/science.aax3072

Beattie, R., & Hippenmeyer, S. (2017). Mechanisms of radial glia progenitor cell lineage progression. *FEBS Letters*, *591*(24), 3993–4008. https://doi.org/10.1002/1873-3468.12906

Begbie, J. (2013a). Induction and patterning of neural crest and ectodermal placodes and their derivatives. *Comprehensive Developmental Neuroscience: Patterning and Cell Type Specification in the Developing CNS and PNS*, 239–258. https://books.google.com/books?hl=en&lr=&id=QvjurYuh-jMC&oi=fnd&pg=PA239&dq=Induction+and+Patterning+of+Neural+Crest+and+Ectodermal+Placodes+and+their+Derivatives&ots=sIihRZw8v5&sig=A8hJKb9D9CA1RW8umOnhtyJ22Ps

Begbie, J. (2013b). Induction and patterning of neural crest and ectodermal placodes and their derivatives. In *Patterning and Cell Type Specification in the Developing CNS and PNS* (pp. 239–258). Elsevier. https://doi.org/10.1016/b978-0-12-397265-1.00212-4

Behesti, H., Holt, J. K. L., & Sowden, J. C. (2006). The level of BMP4 signaling is critical for the regulation of distinct T-box gene expression domains and growth along the dorso-ventral axis of the optic cup. *BMC Developmental Biology*, *6*, 62. https://doi.org/10.1186/1471-213X-6-62

Bendall, S. C., Davis, K. L., Amir, E.-A. D., Tadmor, M. D., Simonds, E. F., Chen, T. J., Shenfeld, D. K., Nolan, G. P., & Pe'er, D. (2014). Single-cell trajectory detection uncovers progression and regulatory coordination in human B cell development. *Cell*, *157*(3), 714–725. https://doi.org/10.1016/j.cell.2014.04.005

Berchtold, M. W., & Villalobo, A. (2014). The many faces of calmodulin in cell proliferation, programmed cell death, autophagy, and cancer. *Biochimica et Biophysica Acta*, *1843*(2), 398–435. https://doi.org/10.1016/j.bbamcr.2013.10.021

Bergen, V., Lange, M., Peidli, S., Wolf, F. A., & Theis, F. J. (2020). Generalizing RNA velocity to transient cell states through dynamical modeling. *Nature Biotechnology*. https://doi.org/10.1038/s41587-020-0591-3

Bergen, V., Soldatov, R. A., Kharchenko, P. V., & Theis, F. J. (2021). RNA velocity—current challenges and future perspectives. *Molecular Systems Biology*, *17*(8), e10282. https://doi.org/10.15252/msb.202110282

Bergmann, S., Ihmels, J., & Barkai, N. (2003). Iterative signature algorithm for the analysis of large-scale gene expression data. *Physical Review. E, Statistical, Nonlinear, and Soft Matter Physics*, *67*(3 Pt 1), 031902. https://doi.org/10.1103/PhysRevE.67.031902

Bertels, S., Jaggy, M., Richter, B., Keppler, S., Weber, K., Genthner, E., Fischer, A. C., Thiel, M., Wegener, M., Greiner, A. M., Autenrieth, T. J., & Bastmeyer, M. (2021). Geometrically defined environments direct cell division rate and subcellular YAP localization in single mouse embryonic stem cells. *Scientific Reports*, *11*(1), 9269. https://doi.org/10.1038/s41598-021-88336-y

Bhaduri, A., Neumann, E. K., Kriegstein, A. R., & Sweedler, J. V. (2021). Identification of lipid heterogeneity and diversity in the developing human brain. *JACS Au*, *1*(12), 2261–2270. https://doi.org/10.1021/jacsau.1c00393

Bhandari, D. R., Coliva, G., Fedorova, M., & Spengler, B. (2020). Single cell analysis by high-resolution atmospheric-pressure MALDI MS imaging. *Methods in Molecular Biology (Clifton, N.J.)*, *2064*, 103–111. https://doi.org/10.1007/978-1-4939-9831-9_8

Bhatia, B., Singhal, S., Jayaram, H., Khaw, P. T., & Limb, G. A. (2010). Adult retinal stem cells revisited. *The Open Ophthalmology Journal*, *4*, 30–38. https://doi.org/10.2174/1874364101004010030

Bingham, E., Chen, J. P., Jankowiak, M., Obermeyer, F., Singh, R., Szerlip, P., Horsfall, P., Goodman, N. D., Pradhan, N., & Karaletsos, T. (2018). *Pyro: Deep universal probabilistic programming*. Retrieved October 14, 2022, from https://www.jmlr.org/papers/volume20/18-403/18-403.pdf

Blanpain, C., & Simons, B. D. (2013). Unravelling stem cell dynamics by lineage tracing. *Nature Reviews. Molecular Cell Biology*, *14*(8), 489–502. https://doi.org/10.1038/nrm3625

Blixt, Å., Mahlapuu, M., Aitola, M., Pelto-Huikko, M., Enerbäck, S., & Carlsson, P. (2000). A forkhead gene, FoxE3, is essential for lens epithelial proliferation and closure of the lens vesicle. *Genes & Development*, *14*(2), 245–254. https://doi.org/10.1101/gad.14.2.245

Bock, C., Datlinger, P., Chardon, F., Coelho, M. A., Dong, M. B., Lawson, K. A., Lu, T., Maroc, L., Norman, T. M., Song, B., Stanley, G., Chen, S., Garnett, M., Li, W., Moffat, J., Qi, L. S., Shapiro, R. S., Shendure, J., Weissman, J. S., & Zhuang, X. (2022). High-content CRISPR screening. *Nature Reviews. Methods Primers*, *2*(1). https://doi.org/10.1038/s43586-022-00098-7

Boles, N. C., Fernandes, M., Swigut, T., Srinivasan, R., Schiff, L., Rada-Iglesias, A., Wang, Q., Saini, J. S., Kiehl, T., Stern, J. H., Wysocka, J., Blenkinsop, T. A., & Temple, S. (2020). Epigenomic and Transcriptomic Changes During Human RPE EMT in a Stem Cell Model of Epiretinal Membrane Pathogenesis and Prevention by Nicotinamide. In *Stem Cell Reports* (Vol. 14, Issue 4, pp. 631–647). https://doi.org/10.1016/j.stemcr.2020.03.009

Bosze, B., Hufnagel, R. B., & Brown, N. L. (2020). Chapter 21 - Specification of retinal cell types. In J. Rubenstein, P. Rakic, B. Chen, & K. Y. Kwan (Eds.), *Patterning and Cell Type Specification in the Developing CNS and PNS (Second Edition)* (pp. 481–504). Academic Press. https://doi.org/10.1016/B978-0-12-814405-3.00021-7

Bouland, G. A., Mahfouz, A., & Reinders, M. J. T. (2021). Differential analysis of binarized single-cell RNA sequencing data captures biological variation. *NAR Genomics and Bioinformatics*, *3*(4), lqab118. https://doi.org/10.1093/nargab/lqab118

Brackston, R. D., Lakatos, E., & Stumpf, M. P. H. (2018). Transition state characteristics during cell differentiation. *PLoS Computational Biology*, *14*(9), e1006405. https://doi.org/10.1371/journal.pcbi.1006405

Bragantini, J., Lange, M., & Royer, L. (2023). Large-scale multi-hypotheses cell tracking using ultrametric contours maps. In *arXiv [cs.CV]*. arXiv. http://arxiv.org/abs/2308.04526

Brandl, C., Zimmermann, S. J., Milenkovic, V. M., Rosendahl, S. M. G., Grassmann, F., Milenkovic, A., Hehr, U., Federlin, M., Wetzel, C. H., Helbig, H., & Weber, B. H. F. (2014). In-depth characterisation of Retinal Pigment Epithelium (RPE) cells derived from human induced pluripotent stem cells (hiPSC). *Neuromolecular Medicine*, *16*(3), 551–564. https://doi.org/10.1007/s12017-014-8308-8

Braun, E., Danan-Gotthold, M., Borm, L. E., Lee, K. W., Vinsland, E., Lönnerberg, P., Hu, L., Li, X., He, X., Andrusivová, Ž., Lundeberg, J., Barker, R. A., Arenas, E., Sundström, E., & Linnarsson, S. (2023). Comprehensive cell atlas of the first-trimester developing human brain. *Science*, *382*(6667), eadf1226. https://doi.org/10.1126/science.adf1226

Brodie-Kommit, J., Clark, B. S., Shi, Q., Shiau, F., Kim, D. W., Langel, J., Sheely, C., Ruzycki, P. A., Fries, M., Javed, A., Cayouette, M., Schmidt, T., Badea, T., Glaser, T., Zhao, H., Singer, J., Blackshaw, S., & Hattar, S. (2021). Atoh7-independent specification of retinal ganglion cell identity. *Science Advances*, *7*(11). https://doi.org/10.1126/sciadv.abe4983

Brunello, L. (2022). Genome-scale single-cell CRISPR screens [Review of *Genome-scale single-cell CRISPR screens*]. *Nature Reviews. Genetics*, *23*(8), 459. https://doi.org/10.1038/s41576-022-00517-1

Buenrostro, J. D., Wu, B., Litzenburger, U. M., Ruff, D., Gonzales, M. L., Snyder, M. P., Chang, H. Y., & Greenleaf, W. J. (2015). Single-cell chromatin accessibility reveals principles of regulatory variation. *Nature*, *523*(7561), 486–490. https://doi.org/10.1038/nature14590

Buettner, F., Natarajan, K. N., Casale, F. P., Proserpio, V., Scialdone, A., Theis, F. J., Teichmann, S. A., Marioni, J. C., & Stegle, O. (2015). Computational analysis of cell-to-cell heterogeneity in single-cell RNA-sequencing data reveals hidden subpopulations of cells. *Nature Biotechnology*, *33*(2), 155–160. https://doi.org/10.1038/nbt.3102

Burbridge, E., & Adrain, C. (2022). Organelle homeostasis: from cellular mechanisms to disease. *The FEBS Journal*, *289*(22), 6822–6831. https://doi.org/10.1111/febs.16667

Burdziak, C., Zhao, C. J., Haviv, D., Alonso-Curbelo, D., Lowe, S. W., & Pe'er, D. (2023). scKINETICS: inference of regulatory velocity with single-cell transcriptomics data. *Bioinformatics* , *39*(39 Suppl 1), i394–i403. https://doi.org/10.1093/bioinformatics/btad267

Butler, A., Hoffman, P., Smibert, P., Papalexi, E., & Satija, R. (2018). Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nature Biotechnology*, *36*(5), 411–420. https://doi.org/10.1038/nbt.4096

Cajal, M., Lawson, K. A., Hill, B., Moreau, A., Rao, J., Ross, A., Collignon, J., & Camus, A. (2012). Clonal and molecular analysis of the prospective anterior neural boundary in the mouse embryo. *Development* , *139*(2), 423–436. https://doi.org/10.1242/dev.075499

Cao, J., Cusanovich, D. A., Ramani, V., Aghamirzaie, D., Pliner, H. A., Hill, A. J., Daza, R. M., McFaline-Figueroa, J. L., Packer, J. S., Christiansen, L., Steemers, F. J., Adey, A. C., Trapnell, C., & Shendure, J. (2018). Joint profiling of chromatin accessibility and gene expression in thousands of single cells. *Science (New York, N.Y.)*, *361*(6409), 1380–1385. https://doi.org/10.1126/science.aau0730

Capolupo, L. (2022). *Digital Phase Contrast on Primary Dermal Human Fibroblasts cells* (Version v0) [Data set]. Zenodo. https://doi.org/10.5281/ZENODO.5996883

Capolupo, L., Burri, O., & Guiet, R. (2022). *Cellpose model for Digital Phase Contrast images* (Version v0) [Data set]. Zenodo. https://doi.org/10.5281/ZENODO.6023317

Capolupo, L., Khven, I., Lederer, A. R., Mazzeo, L., Glousker, G., Ho, S., Russo, F., Montoya, J. P., Bhandari, D. R., Bowman, A. P., Ellis, S. R., Guiet, R., Burri, O., Detzner, J., Muthing, J., Homicsko, K., Kuonen, F., Gilliet, M., Spengler, B., … D'Angelo, G. (2022). Sphingolipids control dermal fibroblast heterogeneity. *Science*, *376*(6590), eabh1623. https://doi.org/10.1126/science.abh1623

Chan, M. M., Smith, Z. D., Grosswendt, S., Kretzmer, H., Norman, T. M., Adamson, B., Jost, M., Quinn, J. J., Yang, D., Jones, M. G., Khodaverdian, A., Yosef, N., Meissner, A., & Weissman, J. S. (2019). Molecular recording of mammalian embryogenesis. *Nature*, *570*(7759), 77–82. https://doi.org/10.1038/s41586-019-1184-5

Chari, T., & Pachter, L. (2023). The specious art of single-cell genomics. *PLoS Computational Biology*, *19*(8), e1011288. https://doi.org/10.1371/journal.pcbi.1011288

Chen, C., Liao, Y., & Peng, G. (2022). Connecting past and present: single-cell lineage tracing. *Protein & Cell*, *13*(11), 790–807. https://doi.org/10.1007/s13238-022-00913-7

Chen, J., Tambalo, M., Barembaum, M., Ranganathan, R., Simões-Costa, M., Bronner, M. E., & Streit, A. (2017). A systems-level approach reveals new gene regulatory modules in the developing ear. *Development* , *144*(8), 1531–1543. https://doi.org/10.1242/dev.148494

Chen, Siyuan, Jiang, W., Du, Y., Yang, M., Pan, Y., Li, H., & Cui, M. (2023). Single-cell analysis technologies for cancer research: from tumor-specific single cell discovery to cancer therapy. *Frontiers in Genetics*, *14*, 1276959. https://doi.org/10.3389/fgene.2023.1276959

Chen, Song, Lake, B. B., & Zhang, K. (2019). High-throughput sequencing of the transcriptome and chromatin accessibility in the same cell. *Nature Biotechnology*, *37*(12), 1452–1457. https://doi.org/10.1038/s41587-019-0290-0

Chen, W., Guillaume-Gentil, O., Rainer, P. Y., Gäbelein, C. G., Saelens, W., Gardeux, V., Klaeger, A., Dainese, R., Zachara, M., Zambelli, T., Vorholt, J. A., & Deplancke, B. (2022). Live-seq enables temporal transcriptomic recording of single cells. *Nature*, *608*(7924), 733–740. https://doi.org/10.1038/s41586-022-05046-9

Chen, X., Xu, N., Li, J., Zhao, M., & Huang, L. (2023). Stem cell therapy for inherited retinal diseases: a systematic review and meta-analysis. *Stem Cell Research & Therapy*, *14*(1). https://doi.org/10.1186/s13287-023-03526-x

Chen, Z., King, W. C., Hwang, A., Gerstein, M., & Zhang, J. (2022). *DeepVelo*: Single-cell transcriptomic deep velocity field learning with neural ordinary differential equations. *Science Advances*, *8*(48), eabq3745. https://doi.org/10.1126/sciadv.abq3745

Cheng, S., Pei, Y., He, L., Peng, G., Reinius, B., Tam, P. P. L., Jing, N., & Deng, Q. (2019). Single-cell RNA-seq reveals cellular heterogeneity of pluripotency transition and X chromosome dynamics during early mouse development. *Cell Reports*, *26*(10), 2593-2607.e3. https://doi.org/10.1016/j.celrep.2019.02.031

Chervov, A., & Zinovyev, A. (2022). Computational challenges of cell cycle analysis using single cell transcriptomics. In *arXiv [q-bio.QM]*. arXiv. http://arxiv.org/abs/2208.05229

Choudhary, P., Booth, H., Gutteridge, A., Surmacz, B., Louca, I., Steer, J., Kerby, J., & Whiting, P. J. (2017). Directing Differentiation of Pluripotent Stem Cells Toward Retinal Pigment Epithelium Lineage. *Stem Cells Translational Medicine*, *6*(2), 490–501. https://doi.org/10.5966/sctm.2016-0088

Choudhary, P., & Whiting, P. J. (2016). A strategy to ensure safety of stem cell-derived retinal pigment epithelium cells. *Stem Cell Research & Therapy*, *7*(1), 127. https://doi.org/10.1186/s13287-016-0380-6

Chow, K.-H. K., Budde, M. W., Granados, A. A., Cabrera, M., Yoon, S., Cho, S., Huang, T.-H., Koulena, N., Frieda, K. L., Cai, L., Lois, C., & Elowitz, M. B. (2021). Imaging cell lineage with a synthetic digital recording system. *Science (New York, N.Y.)*, *372*(6538). https://doi.org/10.1126/science.abb3099

Chu, B. K., Tse, M. J., Sato, R. R., & Read, E. L. (2017). Markov State Models of gene regulatory networks. *BMC Systems Biology*, *11*(1). https://doi.org/10.1186/s12918-017-0394-4

Cliff, T. S., & Dalton, S. (2017). Metabolic switching and cell fate decisions: implications for pluripotency, reprogramming and development. *Current Opinion in Genetics & Development*, *46*, 44–49. https://doi.org/10.1016/j.gde.2017.06.008

Cohen-Salmon, M., El-Amraoui, A., Leibovici, M., & Petit, C. (1997). Otogelin: a glycoprotein specific to the acellular membranes of the inner ear. *Proceedings of the National Academy of Sciences of the United States of America*, *94*(26), 14450–14455. https://doi.org/10.1073/pnas.94.26.14450

Collin, J., Queen, R., Zerti, D., Dorgau, B., Hussain, R., Coxhead, J., Cockell, S., & Lako, M. (2019). Deconstructing retinal organoids: Single cell RNA-Seq reveals the cellular components of human pluripotent stem cell-derived retina. *Stem Cells*, *37*(5), 593–598. https://doi.org/10.1002/stem.2963

Cowan, C. S., Renner, M., De Gennaro, M., Gross-Scherf, B., Goldblum, D., Hou, Y., Munz, M., Rodrigues, T. M., Krol, J., Szikra, T., Cuttat, R., Waldt, A., Papasaikas, P., Diggelmann, R., Patino-Alvarez, C. P., Galliker, P., Spirig, S. E., Pavlinic, D., Gerber-Hollbach, N., … Roska, B. (2020). Cell Types of the Human Retina and Its Organoids at Single-Cell Resolution. *Cell*, *182*(6), 1623-1640.e34. https://doi.org/10.1016/j.cell.2020.08.013

Crespo-Enriquez, I., Partanen, J., Martinez, S., & Echevarria, D. (2012). Fgf8-Related Secondary Organizers Exert Different Polarizing Planar Instructions along the Mouse Anterior Neural Tube. *PloS One*, *7*(7), e39977. https://doi.org/10.1371/journal.pone.0039977

Crick, F. (1966). *Of Molecules and Men*. University of Washington Press. https://philpapers.org/rec/CRIOMA

Cui, H., Maan, H., Vladoiu, M. C., Zhang, J., Taylor, M. D., & Wang, B. (2024). DeepVelo: deep learning extends RNA velocity to multi-lineage systems with cell-specific kinetics. *Genome Biology*, *25*(1). https://doi.org/10.1186/s13059-023-03148-9

Cuomo, A. S. E., Seaton, D. D., McCarthy, D. J., Martinez, I., Bonder, M. J., Garcia-Bernardo, J., Amatya, S., Madrigal, P., Isaacson, A., Buettner, F., Knights, A., Natarajan, K. N., HipSci Consortium, Vallier, L., Marioni, J. C., Chhatriwala, M., & Stegle, O. (2020). Single-cell RNA-sequencing of differentiating iPS cells reveals dynamic genetic effects on gene expression. *Nature Communications*, *11*(1), 810. https://doi.org/10.1038/s41467-020-14457-z

Cvekl, A., & Wang, W.-L. (2009). Retinoic acid signaling in mammalian eye development. *Experimental Eye Research*, *89*(3), 280–291. https://doi.org/10.1016/j.exer.2009.04.012

da Cruz, L., Fynes, K., Georgiadis, O., Kerby, J., Luo, Y. H., Ahmado, A., Vernon, A., Daniels, J. T., Nommiste, B., Hasan, S. M., Gooljar, S. B., Carr, A.-J. F., Vugler, A., Ramsden, C. M., Bictash, M., Fenster, M., Steer, J., Harbinson, T., Wilbrey, A., … Coffey, P. J. (2018). Phase 1 clinical study of an embryonic stem cell–derived retinal pigment epithelium patch in age-related macular degeneration. *Nature Biotechnology*, *36*(4), 328–337. https://doi.org/10.1038/nbt.4114

D'Angelo, G., & La Manno, G. (2023). The lipotype hypothesis. *Nature Reviews. Molecular Cell Biology*, *24*(1), 1–2. https://doi.org/10.1038/s41580-022-00556-w

David, F. P. A., Litovchenko, M., Deplancke, B., & Gardeux, V. (2020). ASAP 2020 update: an open, scalable and interactive web-based portal for (single-cell) omics analyses. *Nucleic Acids Research*, *48*(W1), W403–W414. https://doi.org/10.1093/nar/gkaa412

Davis, D. M., & Dyer, M. A. (2010). Retinal progenitor cells, differentiation, and barriers to cell cycle reentry. *Current Topics in Developmental Biology*, *93*, 175–188. https://doi.org/10.1016/B978-0-12-385044-7.00006-0

De Robertis, E. M. (2006). Spemann's organizer and self-regulation in amphibian embryos. *Nature Reviews. Molecular Cell Biology*, *7*(4), 296–302. https://doi.org/10.1038/nrm1855

Deconinck, L., Cannoodt, R., Saelens, W., Deplancke, B., & Saeys, Y. (2021). Recent advances in trajectory inference from single-cell omics data. *Current Opinion in Systems Biology*, *27*(100344), 100344. https://doi.org/10.1016/j.coisb.2021.05.005

Dehghan, S., Mirshahi, R., Shoae-Hassani, A., & Naseripour, M. (2022). Human-induced pluripotent stem cells-derived retinal pigmented epithelium, a new horizon for cells-based therapies for age-related macular degeneration. *Stem Cell Research & Therapy*, *13*(1). https://doi.org/10.1186/s13287-022-02894-0

Denz, M., Chiantia, S., Herrmann, A., Mueller, P., Korte, T., & Schwarzer, R. (2017). Cell cycle dependent changes in the plasma membrane organization of mammalian cells. *Biochimica et Biophysica Acta*, *1859*(3), 350–359. https://doi.org/10.1016/j.bbamem.2016.12.004

Di Bella, D. J., Habibi, E., Stickels, R. R., Scalia, G., Brown, J., Yadollahpour, P., Yang, S. M., Abbate, C., Biancalani, T., Macosko, E. Z., Chen, F., Regev, A., & Arlotta, P. (2021). Molecular logic of cellular diversification in the mouse cerebral cortex. *Nature*, *595*(7868), 554–559. https://doi.org/10.1038/s41586-021-03670-5

Ding, J., Sharon, N., & Bar-Joseph, Z. (2022). Temporal modelling using single-cell transcriptomics. *Nature Reviews. Genetics*, *23*(6), 355–368. https://doi.org/10.1038/s41576-021-00444-7

Dixit, A., Parnas, O., Li, B., Chen, J., Fulco, C. P., Jerby-Arnon, L., Marjanovic, N. D., Dionne, D., Burks, T., Raychowdhury, R., Adamson, B., Norman, T. M., Lander, E. S., Weissman, J. S., Friedman, N., & Regev, A. (2016). Perturb-Seq: Dissecting Molecular Circuits with Scalable Single-Cell RNA Profiling of Pooled Genetic Screens. *Cell*, *167*(7), 1853-1866.e17. https://doi.org/10.1016/j.cell.2016.11.038

Driskell, R. R., & Watt, F. M. (2015). Understanding fibroblast heterogeneity in the skin. *Trends in Cell Biology*, *25*(2), 92–99. https://doi.org/10.1016/j.tcb.2014.10.001

Dueñas, M. E., Essner, J. J., & Lee, Y. J. (2017). 3D MALDI mass spectrometry imaging of a single cell: Spatial mapping of lipids in the embryonic development of zebrafish. *Scientific Reports*, *7*(1), 14946. https://doi.org/10.1038/s41598-017-14949-x

Eagleson, G., Ferreiro, B., & Harris, W. A. (1995). Fate of the anterior neural ridge and the morphogenesis of the Xenopus forebrain. *Journal of Neurobiology*, *28*(2), 146–158. https://doi.org/10.1002/neu.480280203

Eastman, A. E., Chen, X., Hu, X., Hartman, A. A., Pearlman Morales, A. M., Yang, C., Lu, J., Kueh, H. Y., & Guo, S. (2020). Resolving Cell Cycle Speed in One Snapshot with a Live-Cell Fluorescent Reporter. *Cell Reports*, *31*(12), 107804. https://doi.org/10.1016/j.celrep.2020.107804

Eastman, A. E., & Guo, S. (2020). The palette of techniques for cell cycle analysis. *FEBS Letters*. https://doi.org/10.1002/1873-3468.13842

Eisen, M. B., Spellman, P. T., Brown, P. O., & Botstein, D. (1998). Cluster analysis and display of genome-wide expression patterns. *Proceedings of the National Academy of Sciences of the United States of America*, *95*(25), 14863–14868. https://doi.org/10.1073/pnas.95.25.14863

Erhard, F., Baptista, M. A. P., Krammer, T., Hennig, T., Lange, M., Arampatzi, P., Jürges, C. S., Theis, F. J., Saliba, A.-E., & Dölken, L. (2019). scSLAM-seq reveals core features of transcription dynamics in single cells. *Nature*, *571*(7765), 419–423. https://doi.org/10.1038/s41586-019-1369-y

Ershov, D., Phan, M.-S., Pylvänäinen, J. W., Rigaud, S. U., Le Blanc, L., Charles-Orszag, A., Conway, J. R. W., Laine, R. F., Roy, N. H., Bonazzi, D., Duménil, G., Jacquemet, G., & Tinevez, J.-Y. (2022). TrackMate 7: integrating state-of-the-art segmentation algorithms into tracking pipelines. *Nature Methods*, *19*(7), 829–832. https://doi.org/10.1038/s41592-022-01507-1

Ferrell, J. E., Jr. (2012). Bistability, bifurcations, and waddington's epigenetic landscape. *Current Biology: CB*, *22*(11), R458–R466. https://doi.org/10.1016/j.cub.2012.03.045

Ferry, G. (2019). The structure of DNA. *Nature*, *575*(7781), 35–36. https://doi.org/10.1038/d41586-019-02554-z

Firulli, B. A., Fuchs, R. K., Vincentz, J. W., Clouthier, D. E., & Firulli, A. B. (2014). Hand1 phosphoregulation within the distal arch neural crest is essential for craniofacial morphogenesis. *Development* , *141*(15), 3050–3061. https://doi.org/10.1242/dev.107680

Fortress, A. M., Miyagishima, K. J., Reed, A. A., Temple, S., Clegg, D. O., Tucker, B. A., Blenkinsop, T. A., Harb, G., Greenwell, T. N., Ludwig, T. E., & Bharti, K. (2023). Stem cell sources and characterization in the development of cell-based products for treating retinal disease: An NEI Town Hall report. *Stem Cell Research & Therapy*, *14*(1). https://doi.org/10.1186/s13287-023-03282-y

Frechin, M., Stoeger, T., Daetwyler, S., Gehin, C., Battich, N., Damm, E.-M., Stergiou, L., Riezman, H., & Pelkmans, L. (2015). Cell-intrinsic adaptation of lipid composition to local crowding drives social behaviour. *Nature*, *523*(7558), 88–91. https://doi.org/10.1038/nature14429

Frieda, K. L., Linton, J. M., Hormoz, S., Choi, J., Chow, K.-H. K., Singer, Z. S., Budde, M. W., Elowitz, M. B., & Cai, L. (2017). Synthetic recording and in situ readout of lineage information in single cells. *Nature*, *541*(7635), 107–111. https://doi.org/10.1038/nature20777

Fuhrmann, S., Levine, E. M., & Reh, T. A. (2000). Extraocular mesenchyme patterns the optic vesicle during early eye development in the embryonic chick. *Development* , *127*(21), 4599–4609. https://www.ncbi.nlm.nih.gov/pubmed/11023863

Fuhrmann, Sabine. (2010). Eye morphogenesis and patterning of the optic vesicle. *Current Topics in Developmental Biology*, *93*, 61–84. https://doi.org/10.1016/B978-0-12-385044-7.00003-5

Fuhrmann, Sabine, Zou, C., & Levine, E. M. (2014). Retinal pigment epithelium development, plasticity, and tissue homeostasis. *Experimental Eye Research*, *123*, 141–150. https://doi.org/10.1016/j.exer.2013.09.003

Fujimura, N. (2016). WNT/β-Catenin Signaling in Vertebrate Eye Development. *Frontiers in Cell and Developmental Biology*, *4*, 138. https://doi.org/10.3389/fcell.2016.00138

Gao, M., Qiao, C., & Huang, Y. (2022). UniTVelo: temporally unified RNA velocity reinforces single-cell trajectory inference. *Nature Communications*, *13*(1), 6586. https://doi.org/10.1038/s41467-022-34188-7

Gao, P., Postiglione, M. P., Krieger, T. G., Hernandez, L., Wang, C., Han, Z., Streicher, C., Papusheva, E., Insolera, R., Chugh, K., Kodish, O., Huang, K., Simons, B. D., Luo, L., Hippenmeyer, S., & Shi, S.-H. (2014). Deterministic progenitor behavior and unitary production of neurons in the neocortex. *Cell*, *159*(4), 775–788. https://doi.org/10.1016/j.cell.2014.10.027

Garcia-Ojalvo, J., & Bulut-Karslioglu, A. (2023). On time: developmental timing within and across species. *Development (Cambridge, England)*, *150*(14). https://doi.org/10.1242/dev.201045

Gayoso, A., Weiler, P., Lotfollahi, M., Klein, D., Hong, J., Streets, A., Theis, F. J., & Yosef, N. (2023). Deep generative modeling of transcriptional dynamics for RNA velocity analysis in single cells. *Nature Methods*. https://doi.org/10.1038/s41592-023-01994-w

Gehrs, K. M., Anderson, D. H., Johnson, L. V., & Hageman, G. S. (2006). Age-related macular degeneration--emerging pathogenetic and therapeutic concepts. *Annals of Medicine*, *38*(7), 450–471. https://doi.org/10.1080/07853890600946724

Giaever, G., & Nislow, C. (2014). The yeast deletion collection: A decade of functional genomics. *Genetics*, *197*(2), 451–465. https://doi.org/10.1534/genetics.114.161620

Gilbert, S. F. (2007). *Fate Maps, Gene Expression Maps, And The Evidentiary Structure Of Evolutionary Developmental Biology*. Swarthmore College. https://works.swarthmore.edu/fac-biology/440

Ginhoux, F., Yalin, A., Dutertre, C. A., & Amit, I. (2022). Single-cell immunology: Past, present, and future. *Immunity*, *55*(3), 393–404. https://doi.org/10.1016/j.immuni.2022.02.006

Gitton, Y., Benouaiche, L., Vincent, C., Heude, E., Soulika, M., Bouhali, K., Couly, G., & Levi, G. (2011). Dlx5 and Dlx6 expression in the anterior neural fold is essential for patterning the dorsal nasal capsule. *Development* , *138*(5), 897–903. https://doi.org/10.1242/dev.057505

Gorin, G., Fang, M., Chari, T., & Pachter, L. (2022). RNA velocity unraveled. *PLoS Computational Biology*, *18*(9), e1010492. https://doi.org/10.1371/journal.pcbi.1010492

Gorin, G., Svensson, V., & Pachter, L. (2020). Protein velocity and acceleration from single-cell multiomics experiments. *Genome Biology*, *21*(1), 39. https://doi.org/10.1186/s13059-020-1945-3

Gorin, G., Vastola, J. J., & Pachter, L. (2023). Studying stochastic systems biology of the cell with single-cell genomics data. *Cell Systems*, *14*(10), 822-843.e22. https://doi.org/10.1016/j.cels.2023.08.004

Griffith, J. S. (1968). Mathematics of cellular control processes. II. Positive feedback to one gene. *Journal of Theoretical Biology*, *20*(2), 209–216. https://doi.org/10.1016/0022-5193(68)90190-2

Grove, E. A., & Monuki, E. S. (2020). Chapter 1 - Morphogens, patterning centers, and their mechanisms of action. In J. Rubenstein, P. Rakic, B. Chen, & K. Y. Kwan (Eds.), *Patterning and Cell Type Specification in the Developing CNS and PNS (Second Edition)* (pp. 3–21). Academic Press. https://doi.org/10.1016/B978-0-12-814405-3.00001-1

Gruber, J. J., Geller, B., Lipchik, A. M., Chen, J., Salahudeen, A. A., Ram, A. N., Ford, J. M., Kuo, C. J., & Snyder, M. P. (2019). HAT1 Coordinates Histone Production and Acetylation via H4 Promoter Binding. *Molecular Cell*, *75*(4), 711-724.e5. https://doi.org/10.1016/j.molcel.2019.05.034

Grün, D. (2020). Revealing dynamics of gene expression variability in cell state space. *Nature Methods*, *17*(1), 45–49. https://doi.org/10.1038/s41592-019-0632-3

Gu, Y., Blaauw, D., & Welch, J. D. (2022). Bayesian inference of rna velocity from multi-lineage single-cell data. *BioRxiv*. https://www.biorxiv.org/content/10.1101/2022.07.08.499381.abstract

Gunawan, I., Vafaee, F., Meijering, E., & Lock, J. G. (2023). An introduction to representation learning for single-cell data analysis. *Cell Reports Methods*, *3*(8), 100547. https://doi.org/10.1016/j.crmeth.2023.100547

Guo, C., Kong, W., Kamimoto, K., Rivera-Gonzalez, G. C., Yang, X., Kirita, Y., & Morris, S. A. (2019). CellTag Indexing: Genetic barcode-based sample multiplexing for single-cell genomics. *Genome Biology*, *20*(1), 1–13. https://doi.org/10.1186/s13059-019-1699-y

Gupta, P. B., Fillmore, C. M., Jiang, G., Shapira, S. D., Tao, K., Kuperwasser, C., & Lander, E. S. (2011). Stochastic state transitions give rise to phenotypic equilibrium in populations of cancer cells. *Cell*, *146*(4), 633–644. https://doi.org/10.1016/j.cell.2011.07.026

Gupta, S., Lytvynchuk, L., Ardan, T., Studenovska, H., Sharma, R., Faura, G., Eide, L., Shanker Verma, R., Znaor, L., Erceg, S., Stieger, K., Motlik, J., Petrovski, G., & Bharti, K. (2023). Progress in stem cells-based replacement therapy for Retinal pigment epithelium: In vitro differentiation to in vivo delivery. *Stem Cells Translational Medicine*, *12*(8), 536–552. https://doi.org/10.1093/stcltm/szad039

Hadziahmetovic, M., & Malek, G. (2021). Age-related macular degeneration revisited: From pathology and cellular stress to potential therapies. *Frontiers in Cell and Developmental Biology*, *8*. https://doi.org/10.3389/fcell.2020.612812

Hagemann-Jensen, M., Ziegenhain, C., & Sandberg, R. (2022). Scalable single-cell RNA sequencing from full transcripts with Smart-seq3xpress. *Nature Biotechnology*, *40*(10), 1452–1457. https://doi.org/10.1038/s41587-022-01311-4

Haghverdi, L., Büttner, M., Wolf, F. A., Buettner, F., & Theis, F. J. (2016). Diffusion pseudotime robustly reconstructs lineage branching. *Nature Methods*, *13*(10), 845–848. https://doi.org/10.1038/nmeth.3971

Haghverdi, L., & Ludwig, L. S. (2023). Single-cell multi-omics and lineage tracing to dissect cell fate decision-making. *Stem Cell Reports*, *18*(1), 13–25. https://doi.org/10.1016/j.stemcr.2022.12.003

Hahaut, V., Pavlinic, D., Carbone, W., Schuierer, S., Balmer, P., Quinodoz, M., Renner, M., Roma, G., Cowan, C. S., & Picelli, S. (2022). Fast and highly sensitive full-length single-cell RNA sequencing using FLASH-seq. *Nature Biotechnology*, *40*(10), 1447–1451. https://doi.org/10.1038/s41587-022-01312-3

Haniffa, M., Taylor, D., Linnarsson, S., Aronow, B. J., Bader, G. D., Barker, R. A., Camara, P. G., Camp, J. G., Chédotal, A., Copp, A., Etchevers, H. C., Giacobini, P., Göttgens, B., Guo, G., Hupalowska, A., James, K. R., Kirby, E., Kriegstein, A., Lundeberg, J., … Human Cell Atlas Developmental Biological Network. (2021). A roadmap for the Human Developmental Cell Atlas. *Nature*, *597*(7875), 196–205. https://doi.org/10.1038/s41586-021-03620-1

Hannun, Y. A., & Obeid, L. M. (2018). Sphingolipids and their metabolism in physiology and disease. *Nature Reviews. Molecular Cell Biology*, *19*(3), 175–191. https://doi.org/10.1038/nrm.2017.107

Haque, A., Engel, J., Teichmann, S. A., & Lönnberg, T. (2017). A practical guide to single-cell RNA-sequencing for biomedical research and clinical applications. *Genome Medicine*, *9*(1). https://doi.org/10.1186/s13073-017-0467-4

Harris, L., Zalucki, O., & Piper, M. (2018). BrdU/EdU dual labeling to determine the cell-cycle dynamics of defined cellular subpopulations. *Journal of Molecular Histology*, *49*(3), 229–234. https://doi.org/10.1007/s10735-018-9761-8

Hashimshony, T., Wagner, F., Sher, N., & Yanai, I. (2012). CEL-Seq: single-cell RNA-Seq by multiplexed linear amplification. *Cell Reports*, *2*(3), 666–673. https://doi.org/10.1016/j.celrep.2012.08.003

Hazim, R. A., Karumbayaram, S., Jiang, M., Dimashkie, A., Lopes, V. S., Li, D., Burgess, B. L., Vijayaraj, P., Alva-Ornelas, J. A., Zack, J. A., Kohn, D. B., Gomperts, B. N., Pyle, A. D., Lowry, W. E., & Williams, D. S. (2017). Differentiation of RPE cells from integration-free iPS cells and their cell biological characterization. *Stem Cell Research & Therapy*, *8*(1), 217. https://doi.org/10.1186/s13287-017-0652-9

He, D., Zakeri, M., Sarkar, H., Soneson, C., Srivastava, A., & Patro, R. (2022). Alevin-fry unlocks rapid, accurate and memory-frugal quantification of single-cell RNA-seq data. *Nature Methods*, *19*(3), 316–322. https://doi.org/10.1038/s41592-022-01408-3

He, Z., Maynard, A., Jain, A., Gerber, T., Petri, R., Lin, H.-C., Santel, M., Ly, K., Dupré, J.-S., Sidow, L., Sanchis Calleja, F., Jansen, S. M. J., Riesenberg, S., Camp, J. G., & Treutlein, B. (2022). Lineage recording in human cerebral organoids. *Nature Methods*, *19*(1), 90–99. https://doi.org/10.1038/s41592-021-01344-8

Hendriks, G.-J., Jung, L. A., Larsson, A. J. M., Lidschreiber, M., Andersson Forsman, O., Lidschreiber, K., Cramer, P., & Sandberg, R. (2019). NASC-seq monitors RNA synthesis in single cells. *Nature Communications*, *10*(1), 3138. https://doi.org/10.1038/s41467-019-11028-9

Heyningen S Van. (1974). Cholera toxin: interaction of subunits with ganglioside GM1. *Science (New York, N.Y.)*, *183*(4125), 656–657. https://doi.org/10.1126/science.183.4125.656

Higuchi, A., Kumar, S. S., Benelli, G., Alarfaj, A. A., Munusamy, M. A., Umezawa, A., & Murugan, K. (2017). Stem Cell Therapies for Reversing Vision Loss. *Trends in Biotechnology*, *35*(11), 1102–1117. https://doi.org/10.1016/j.tibtech.2017.06.016

Hou, W., Ji, Z., Ji, H., & Hicks, S. C. (2020). A systematic evaluation of single-cell RNA-sequencing imputation methods. *Genome Biology*, *21*(1). https://doi.org/10.1186/s13059-020-02132-x

Hsu, Y.-C. (2015). Theory and practice of lineage tracing. *Stem Cells (Dayton, Ohio)*, *33*(11), 3197–3204. https://doi.org/10.1002/stem.2123

Hu, Y., Wang, X., Hu, B., Mao, Y., Chen, Y., Yan, L., Yong, J., Dong, J., Wei, Y., Wang, W., Wen, L., Qiao, J., & Tang, F. (2019). Dissecting the transcriptome landscape of the human fetal neural retina and retinal

pigment epithelium by single-cell RNA-seq analysis. *PLoS Biology*, *17*(7), e3000365. https://doi.org/10.1371/journal.pbio.3000365

Huang, S. (2012). The molecular and mathematical basis of Waddington's epigenetic landscape: a framework for post-Darwinian biology? *BioEssays: News and Reviews in Molecular, Cellular and Developmental Biology*, *34*(2), 149–157. https://doi.org/10.1002/bies.201100031

Hub Zwart, N. (2007). Genomics and self-knowledge: implications for societal research and debate. *New Genetics and Society*, *26*(2), 181–202. https://doi.org/10.1080/14636770701466881

Imuta, Y., Nishioka, N., Kiyonari, H., & Sasaki, H. (2009). Short limbs, cleft palate, and delayed formation of flat proliferative chondrocytes in mice with targeted disruption of a putative protein kinase gene, Pkdcc (AW548124). *Developmental Dynamics: An Official Publication of the American Association of Anatomists*, *238*(1), 210–222. https://doi.org/10.1002/dvdy.21822

Islam, S., Kjällquist, U., Moliner, A., Zajac, P., Fan, J.-B., Lönnerberg, P., & Linnarsson, S. (2011). Characterization of the single-cell transcriptional landscape by highly multiplex RNA-seq. *Genome Research*, *21*(7), 1160–1167. https://doi.org/10.1101/gr.110882.110

Jacewicz, M., Clausen, H., Nudelman, E., Donohue-Rolfe, A., & Keusch, G. T. (1986). Pathogenesis of shigella diarrhea. XI. Isolation of a shigella toxin-binding glycolipid from rabbit jejunum and HeLa cells and its identification as globotriaosylceramide. *The Journal of Experimental Medicine*, *163*(6), 1391–1404. https://doi.org/10.1084/jem.163.6.1391

Jacob, F., & Monod, J. (1961). Genetic Regulatory Mechanisms in the Synthesis of Proteins. *J. Mol. Biol.*, *3*, 318–356.

Jaeger, J., & Monk, N. (2014). Bioattractors: dynamical systems theory and the evolution of regulatory processes. *The Journal of Physiology*, *592*(11), 2267–2281. https://doi.org/10.1113/jphysiol.2014.272385

Jeon, K. W. (Ed.). (2013). Chapter Seven - Role of Cyclin B1 Levels in DNA Damage and DNA Damage-Induced Senescence. In *International Review of Cell and Molecular Biology. Academic P*.

Jinek, M., Chylinski, K., Fonfara, I., Hauer, M., Doudna, J. A., & Charpentier, E. (2012). A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. *Science (New York, N.Y.)*, *337*(6096), 816–821. https://doi.org/10.1126/science.1225829

Johnson, W. E., Li, C., & Rabinovic, A. (2007). Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics (Oxford, England)*, *8*(1), 118–127. https://doi.org/10.1093/biostatistics/kxj037

Joost, S., Zeisel, A., Jacob, T., Sun, X., La Manno, G., Lönnerberg, P., Linnarsson, S., & Kasper, M. (2016). Single-cell transcriptomics reveals that differentiation and spatial signatures shape epidermal and hair follicle heterogeneity. *Cell Systems*, *3*(3), 221-237.e9. https://doi.org/10.1016/j.cels.2016.08.010

Joshi, R., Mankowski, W., Winter, M., Saini, J. S., Blenkinsop, T. A., Stern, J. H., Temple, S., & Cohen, A. R. (2016). Automated Measurement of Cobblestone Morphology for Characterizing Stem Cell Derived Retinal Pigment Epithelial Cell Cultures. *Journal of Ocular Pharmacology and Therapeutics: The Official Journal of the Association for Ocular Pharmacology and Therapeutics*, *32*(5), 331–339. https://doi.org/10.1089/jop.2015.0163

Jovic, D., Liang, X., Zeng, H., Lin, L., Xu, F., & Luo, Y. (2022). Single-cell RNA sequencing technologies and applications: A brief overview. *Clinical and Translational Medicine*, *12*(3). https://doi.org/10.1002/ctm2.694

Jung, A. R., Jung, C.-H., Noh, J. K., Lee, Y. C., & Eun, Y.-G. (2020). Epithelial-mesenchymal transition gene signature is associated with prognosis and tumor microenvironment in head and neck squamous cell carcinoma. *Scientific Reports*, *10*(1), 3652. https://doi.org/10.1038/s41598-020-60707-x

Kagiyama, Y., Gotouda, N., Sakagami, K., Yasuda, K., Mochii, M., & Araki, M. (2005). Extraocular dorsal signal affects the developmental fate of the optic vesicle and patterns the optic neuroepithelium. *Development, Growth & Differentiation*, *47*(8), 523–536. https://doi.org/10.1111/j.1440-169X.2005.00828.x

Kaminow, B., Yunusov, D., & Dobin, A. (2021). STARsolo: accurate, fast and versatile mapping/quantification of single-cell and single-nucleus RNA-seq data. In *bioRxiv* (p. 2021.05.05.442755). https://doi.org/10.1101/2021.05.05.442755

Kang, R. B., Li, Y., Rosselot, C., Zhang, T., Siddiq, M., Rajbhandari, P., Stewart, A. F., Scott, D. K., Garcia-Ocana, A., & Lu, G. (2023). Single-nucleus RNA sequencing of human pancreatic islets identifies novel gene sets and distinguishes β-cell subpopulations with dynamic transcriptome profiles. *Genome Medicine*, *15*(1). https://doi.org/10.1186/s13073-023-01179-2

Karemaker, I. D., & Vermeulen, M. (2018). Single-cell DNA methylation profiling: Technologies and biological applications. *Trends in Biotechnology*, *36*(9), 952–965. https://doi.org/10.1016/j.tibtech.2018.04.002

Kasberg, A. D., Brunskill, E. W., & Steven Potter, S. (2013). SP8 regulates signaling centers during craniofacial development. *Developmental Biology*, *381*(2), 312–323. https://doi.org/10.1016/j.ydbio.2013.07.007

Kashani, A. H., Lebkowski, J. S., Rahhal, F. M., Avery, R. L., Salehi-Had, H., Dang, W., Lin, C.-M., Mitra, D., Zhu, D., Thomas, B. B., Hikita, S. T., Pennington, B. O., Johnson, L. V., Clegg, D. O., Hinton, D. R., & Humayun, M. S. (2018). A bioengineered retinal pigment epithelial monolayer for advanced, dry age-related macular degeneration. *Science Translational Medicine*, *10*(435). https://doi.org/10.1126/scitranslmed.aao4097

Kaya-Okur, H. S., Wu, S. J., Codomo, C. A., Pledger, E. S., Bryson, T. D., Henikoff, J. G., Ahmad, K., & Henikoff, S. (2019). CUT&Tag for efficient epigenomic profiling of small samples and single cells. *Nature Communications*, *10*(1), 1–10. https://doi.org/10.1038/s41467-019-09982-5

Kebschull, J. M., & Zador, A. M. (2018). Cellular barcoding: lineage tracing, screening and beyond. *Nature Methods*, *15*(11), 871–879. https://doi.org/10.1038/s41592-018-0185-x

Kester, L., & van Oudenaarden, A. (2018). Single-Cell Transcriptomics Meets Lineage Tracing. *Cell Stem Cell*, *23*(2), 166–179. https://doi.org/10.1016/j.stem.2018.04.014

Kiecker, C., Bates, T., & Bell, E. (2016). Molecular specification of germ layers in vertebrate embryos. *Cellular and Molecular Life Sciences: CMLS*, *73*(5), 923–947. https://doi.org/10.1007/s00018-015-2092-y

Kim, N., Kang, H., Jo, A., Yoo, S.-A., & Lee, H.-O. (2023). Perspectives on single-nucleus RNA sequencing in different cell types and tissues. *Journal of Pathology and Translational Medicine*, *57*(1), 52–59. https://doi.org/10.4132/jptm.2022.12.19

Klein, A. M., Mazutis, L., Akartuna, I., Tallapragada, N., Veres, A., Li, V., Peshkin, L., Weitz, D. A., & Kirschner, M. W. (2015). Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. *Cell*, *161*(5), 1187–1201. https://doi.org/10.1016/j.cell.2015.04.044

Klein, A. M., & Simons, B. D. (2011). Universal patterns of stem cell fate in cycling adult tissues. *Development (Cambridge, England)*, *138*(15), 3103–3111. https://doi.org/10.1242/dev.060103

Kompauer, M., Heiles, S., & Spengler, B. (2017). Atmospheric pressure MALDI mass spectrometry imaging of tissues and cells at 1.4-μm lateral resolution. *Nature Methods*, *14*(1), 90–96. https://doi.org/10.1038/nmeth.4071

Kong, W., Biddy, B. A., Kamimoto, K., Amrute, J. M., Butka, E. G., & Morris, S. A. (2020). CellTagging: combinatorial indexing to simultaneously map lineage and identity at single-cell resolution. *Nature Protocols*, *15*(3), 750–772. https://doi.org/10.1038/s41596-019-0247-2

Kramer, B. A., Sarabia del Castillo, J., & Pelkmans, L. (2022). Multimodal perception links cellular state to decision-making in single cells. *Science (New York, N.Y.)*, *377*(6606), 642–648. https://doi.org/10.1126/science.abf4062

Krenning, L., Sonneveld, S., & Tanenbaum, M. E. (2022). Time-resolved single-cell sequencing identifies multiple waves of mRNA decay during the mitosis-to-G1 phase transition. *ELife*, *11*. https://doi.org/10.7554/eLife.71356

Kulkarni, A., Anderson, A. G., Merullo, D. P., & Konopka, G. (2019). Beyond bulk: a review of single cell transcriptomics methodologies and applications. *Current Opinion in Biotechnology*, *58*, 129–136. https://doi.org/10.1016/j.copbio.2019.03.001

Kumamoto, T., & Hanashima, C. (2017). Evolutionary conservation and conversion of Foxg1 function in brain development. *Development, Growth & Differentiation*, *59*(4), 258–269. https://doi.org/10.1111/dgd.12367

Kumar, P., Tan, Y., & Cahan, P. (2017). Understanding development and stem cells using single cell-based analyses of gene expression. *Development* , *144*(1), 17–32. https://doi.org/10.1242/dev.133058

Kwon, H.-J., Bhat, N., Sweet, E. M., Cornell, R. A., & Riley, B. B. (2010). Identification of early requirements for preplacodal ectoderm and sensory organ development. *PLoS Genetics*, *6*(9), e1001133. https://doi.org/10.1371/journal.pgen.1001133

La Manno, G., Gyllborg, D., Codeluppi, S., Nishimura, K., Salto, C., Zeisel, A., Borm, L. E., Stott, S. R. W., Toledo, E. M., Villaescusa, J. C., Lönnerberg, P., Ryge, J., Barker, R. A., Arenas, E., & Linnarsson, S. (2016). Molecular Diversity of Midbrain Development in Mouse, Human, and Stem Cells. *Cell*, *167*(2), 566-580.e19. https://doi.org/10.1016/j.cell.2016.09.027

La Manno, G., Siletti, K., Furlan, A., Gyllborg, D., Vinsland, E., Mossi Albiach, A., Mattsson Langseth, C., Khven, I., Lederer, A. R., Dratva, L. M., Johnsson, A., Nilsson, M., Lönnerberg, P., & Linnarsson, S. (2021).

Molecular architecture of the developing mouse brain. *Nature*, *596*(7870), 92–96. https://doi.org/10.1038/s41586-021-03775-x

La Manno, G., Soldatov, R., Zeisel, A., Braun, E., Hochgerner, H., Petukhov, V., Lidschreiber, K., Kastriti, M. E., Lönnerberg, P., Furlan, A., Fan, J., Borm, L. E., Liu, Z., van Bruggen, D., Guo, J., He, X., Barker, R., Sundström, E., Castelo-Branco, G., … Kharchenko, P. V. (2018). RNA velocity of single cells. *Nature*, *560*(7719), 494–498. https://doi.org/10.1038/s41586-018-0414-6

Labib, M., & Kelley, S. O. (2020). Single-cell analysis targeting the proteome. *Nature Reviews Chemistry*, *4*(3), 143–158. https://doi.org/10.1038/s41570-020-0162-7

Lähnemann, D., Köster, J., Szczurek, E., McCarthy, D. J., Hicks, S. C., Robinson, M. D., Vallejos, C. A., Campbell, K. R., Beerenwinkel, N., Mahfouz, A., Pinello, L., Skums, P., Stamatakis, A., Attolini, C. S.-O., Aparicio, S., Baaijens, J., Balvert, M., Barbanson, B. de, Cappuccio, A., … Schönhuth, A. (2020). Eleven grand challenges in single-cell data science. *Genome Biology*, *21*(1), 31. https://doi.org/10.1186/s13059-020-1926-6

Lange, M., Bergen, V., Klein, M., Setty, M., Reuter, B., Bakhti, M., Lickert, H., Ansari, M., Schniering, J., Schiller, H. B., Pe'er, D., & Theis, F. J. (2022). CellRank for directed single-cell fate mapping. *Nature Methods*, *19*(2), 159–170. https://doi.org/10.1038/s41592-021-01346-6

Lear, S. K., & Shipman, S. L. (2023). Molecular recording: transcriptional data collection into the genome. *Current Opinion in Biotechnology*, *79*(102855), 102855. https://doi.org/10.1016/j.copbio.2022.102855

Lederer, A. R., & La Manno, G. (2020). The emergence and promise of single-cell temporal-omics approaches. *Current Opinion in Biotechnology*, *63*, 70–78. https://doi.org/10.1016/j.copbio.2019.12.005

Lederer, A. R., Leonardi, M., Talamanca, L., Herrera, A., Droin, C., Khven, I., Carvalho, H. J. F., Valente, A., Dominguez Mantes, A., Mulet Arabí, P., Pinello, L., Naef, F., & La Manno, G. (2024). Statistical inference with a manifold-constrained RNA velocity model uncovers cell cycle speed modulations. In *bioRxiv*. https://doi.org/10.1101/2024.01.18.576093

Lee, J., Choi, A., Cho, S.-Y., Jun, Y., Na, D., Lee, A., Jang, G., Kwon, J. Y., Kim, J., Lee, S., & Lee, C. (2021). Genome-scale CRISPR screening identifies cell cycle and protein ubiquitination processes as druggable targets for erlotinib-resistant lung cancer. *Molecular Oncology*, *15*(2), 487–502. https://doi.org/10.1002/1878-0261.12853

Levental, I., & Lyman, E. (2023). Regulation of membrane protein structure and function by their lipid nano-environment. *Nature Reviews. Molecular Cell Biology*, *24*(2), 107–122. https://doi.org/10.1038/s41580-022-00524-4

Lewis, M. (2011). A tale of two repressors. *Journal of Molecular Biology*, *409*(1), 14–27. https://doi.org/10.1016/j.jmb.2011.02.023

Li, C., Virgilio, M. C., Collins, K. L., & Welch, J. D. (2022). Multi-omic single-cell velocity models epigenome–transcriptome interactions and improves cell fate prediction. *Nature Biotechnology*, 1–12. https://doi.org/10.1038/s41587-022-01476-y

Li, J., Pan, X., Yuan, Y., & Shen, H.-B. (2023). TFvelo: gene regulation inspired RNA velocity estimation. In *bioRxiv*. https://doi.org/10.1101/2023.07.12.548785

Li, S., Zhang, P., Chen, W., Ye, L., Brannan, K. W., Le, N.-T., Abe, J.-I., Cooke, J. P., & Wang, G. (2023). A relay velocity model infers cell-dependent RNA velocity. *Nature Biotechnology*, 1–10. https://doi.org/10.1038/s41587-023-01728-5

Liang, Y., Sun, X., Duan, C., Tang, S., & Chen, J. (2023). Application of patient-derived induced pluripotent stem cells and organoids in inherited retinal diseases. *Stem Cell Research & Therapy*, *14*(1). https://doi.org/10.1186/s13287-023-03564-5

Liberali, P., Snijder, B., & Pelkmans, L. (2014). A hierarchical map of regulatory genetic interactions in membrane trafficking. *Cell*, *157*(6), 1473–1487. https://doi.org/10.1016/j.cell.2014.04.029

Lidgerwood, G. E., Senabouth, A., Smith-Anttila, C. J. A., Gnanasambandapillai, V., Kaczorowski, D. C., Amann-Zalcenstein, D., Fletcher, E. L., Naik, S. H., Hewitt, A. W., Powell, J. E., & Pébay, A. (2021). Transcriptomic Profiling of Human Pluripotent Stem Cell-derived Retinal Pigment Epithelium over Time. *Genomics, Proteomics & Bioinformatics*, *19*(2), 223–242. https://doi.org/10.1016/j.gpb.2020.08.002

Liggett, L. A., & Sankaran, V. G. (2020). Unraveling hematopoiesis through the lens of genomics. *Cell*, *182*(6), 1384–1400. https://doi.org/10.1016/j.cell.2020.08.030

Lin, H.-C., He, Z., Ebert, S., Schörnig, M., Santel, M., Nikolova, M. T., Weigert, A., Hevers, W., Kasri, N. N., Taverna, E., Camp, J. G., & Treutlein, B. (2021). NGN2 induces diverse neuron types from human pluripotency. *Stem Cell Reports*, *16*(9), 2118–2127. https://doi.org/10.1016/j.stemcr.2021.07.006

Lindeboom, R. G. H., Regev, A., & Teichmann, S. A. (2021). Towards a Human Cell Atlas: Taking notes from the past. *Trends in Genetics: TIG*, *37*(7), 625–630. https://doi.org/10.1016/j.tig.2021.03.007

Linderman, G. C. (2021). Dimensionality reduction of single-cell RNA-seq data. In *Methods in Molecular Biology* (pp. 331–342). Springer US. https://doi.org/10.1007/978-1-0716-1307-8_18

Lister, R., O'Malley, R. C., Tonti-Filippini, J., Gregory, B. D., Berry, C. C., Millar, A. H., & Ecker, J. R. (2008). Highly integrated single-base resolution maps of the epigenome in Arabidopsis. *Cell*, *133*(3), 523–536. https://doi.org/10.1016/j.cell.2008.03.029

Liu, J., Yang, M., Zhao, W., & Zhou, X. (2022). CCPE: cell cycle pseudotime estimation for single cell RNA-seq data. *Nucleic Acids Research*, *50*(2), 704–716. https://doi.org/10.1093/nar/gkab1236

Lo Giudice, Q., Leleu, M., La Manno, G., & Fabre, P. J. (2019). Single-cell transcriptional logic of cell-fate specification and axon guidance in early-born retinal neurons. *Development* , *146*(17). https://doi.org/10.1242/dev.178103

Lu, Y., Shiau, F., Yi, W., Lu, S., Wu, Q., Pearson, J. D., Kallman, A., Zhong, S., Hoang, T., Zuo, Z., Zhao, F., Zhang, M., Tsai, N., Zhuo, Y., He, S., Zhang, J., Stein-O'Brien, G. L., Sherman, T. D., Duan, X., … Clark, B. S. (2020). Single-Cell Analysis of Human Retina Identifies Evolutionarily Conserved and Species-Specific Mechanisms Controlling Development. *Developmental Cell*, *53*(4), 473-491.e9. https://doi.org/10.1016/j.devcel.2020.04.009

Luecken, M. D., Büttner, M., Chaichoompu, K., Danese, A., Interlandi, M., Mueller, M. F., Strobl, D. C., Zappia, L., Dugas, M., Colomé-Tatché, M., & Theis, F. J. (2022). Benchmarking atlas-level data integration in single-cell genomics. *Nature Methods*, *19*(1), 41–50. https://doi.org/10.1038/s41592-021-01336-8

Luecken, M. D., & Theis, F. J. (2019). Current best practices in single-cell RNA-seq analysis: a tutorial. *Molecular Systems Biology*, *15*(6), e8746. https://doi.org/10.15252/msb.20188746

Lukowski, S. W., Lo, C. Y., Sharov, A. A., Nguyen, Q., Fang, L., Hung, S. S., Zhu, L., Zhang, T., Grünert, U., Nguyen, T., Senabouth, A., Jabbari, J. S., Welby, E., Sowden, J. C., Waugh, H. S., Mackey, A., Pollock, G., Lamb, T. D., Wang, P.-Y., … Wong, R. C. (2019). A single-cell transcriptome atlas of the adult human retina. *The EMBO Journal*, *38*(18), e100811. https://doi.org/10.15252/embj.2018100811

Ma, S., Zhang, B., LaFave, L. M., Earl, A. S., Chiang, Z., Hu, Y., Ding, J., Brack, A., Kartha, V. K., Tay, T., Law, T., Lareau, C., Hsu, Y.-C., Regev, A., & Buenrostro, J. D. (2020). Chromatin potential identified by shared single-cell profiling of RNA and chromatin. *Cell*, *183*(4), 1103-1116.e20. https://doi.org/10.1016/j.cell.2020.09.056

Ma, Y., Kanakousaki, K., & Buttitta, L. (2015). How the cell cycle impacts chromatin architecture and influences cell fate. *Frontiers in Genetics*, *6*. https://doi.org/10.3389/fgene.2015.00019

Ma, Y., McKay, D. J., & Buttitta, L. (2019). Changes in chromatin accessibility ensure robust cell cycle exit in terminally differentiated cells. *PLoS Biology*, *17*(9), e3000378. https://doi.org/10.1371/journal.pbio.3000378

MacLean, A. L., Hong, T., & Nie, Q. (2018). Exploring intermediate cell states through the lens of single cells. *Current Opinion in Systems Biology*, *9*, 32–41. https://doi.org/10.1016/j.coisb.2018.02.009

Macosko, E. Z., Basu, A., Satija, R., Nemesh, J., Shekhar, K., Goldman, M., Tirosh, I., Bialas, A. R., Kamitaki, N., Martersteck, E. M., Trombetta, J. J., Weitz, D. A., Sanes, J. R., Shalek, A. K., Regev, A., & McCarroll, S. A. (2015). Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets. *Cell*, *161*(5), 1202–1214. https://doi.org/10.1016/j.cell.2015.05.002

Maehle, A.-H. (2011). Ambiguous cells: the emergence of the stem cell concept in the nineteenth and twentieth centuries. *Notes and Records of the Royal Society of London*, *65*(4), 359–378. https://doi.org/10.1098/rsnr.2011.0023

Maizels, R. J., Snell, D. M., & Briscoe, J. (2023). Deep dynamical modelling of developmental trajectories with temporal transcriptomics. In *bioRxiv* (p. 2023.07.06.547989). https://doi.org/10.1101/2023.07.06.547989

Mandai, M., Watanabe, A., Kurimoto, Y., Hirami, Y., Morinaga, C., Daimon, T., Fujihara, M., Akimaru, H., Sakai, N., Shibata, Y., Terada, M., Nomiya, Y., Tanishima, S., Nakamura, M., Kamao, H., Sugita, S., Onishi, A., Ito, T., Fujita, K., … Takahashi, M. (2017). Autologous Induced Stem-Cell–Derived Retinal Cells for Macular Degeneration. *The New England Journal of Medicine*, *376*(11), 1038–1046. https://doi.org/10.1056/NEJMoa1608368

Mao, X., An, Q., Xi, H., Yang, X.-J., Zhang, X., Yuan, S., Wang, J., Hu, Y., Liu, Q., & Fan, G. (2019). Single-Cell RNA Sequencing of hESC-Derived 3D Retinal Organoids Reveals Novel Genes Regulating RPC Commitment in Early Human Retinogenesis. *Stem Cell Reports*, *13*(4), 747–760. https://doi.org/10.1016/j.stemcr.2019.08.012

Marquardt, T., Ashery-Padan, R., Andrejewski, N., Scardigli, R., Guillemot, F., & Gruss, P. (2001). Pax6 is required for the multipotent state of retinal progenitor cells. *Cell*, *105*(1), 43–55. https://doi.org/10.1016/s0092-8674(01)00295-1

Marsh, B., & Blelloch, R. (2020). Single nuclei RNA-seq of mouse placental labyrinth development. *ELife*, *9*. https://doi.org/10.7554/eLife.60266

Martínez-Morales, J. R., Dolez, V., Rodrigo, I., Zaccarini, R., Leconte, L., Bovolenta, P., & Saule, S. (2003). OTX2 activates the molecular network underlying retina pigment epithelium differentiation. *The Journal of Biological Chemistry*, *278*(24), 21721–21731. https://doi.org/10.1074/jbc.M301708200

Marx, V. (2023). Method of the year: long-read sequencing. *Nature Methods*, *20*(1), 6–11. https://doi.org/10.1038/s41592-022-01730-w

Matson, J. P., & Cook, J. G. (2017). Cell cycle proliferation decisions: the impact of single cell analyses. *The FEBS Journal*, *284*(3), 362–375. https://doi.org/10.1111/febs.13898

Matthews, H. K., Bertoli, C., & de Bruin, R. A. M. (2022). Cell cycle control in cancer. *Nature Reviews. Molecular Cell Biology*, *23*(1), 74–88. https://doi.org/10.1038/s41580-021-00404-3

McCracken, I. R., Taylor, R. S., Kok, F. O., de la Cuesta, F., Dobie, R., Henderson, B. E. P., Mountford, J. C., Caudrillier, A., Henderson, N. C., Ponting, C. P., & Baker, A. H. (2020). Transcriptional dynamics of pluripotent stem cell-derived endothelial cell differentiation revealed by single-cell RNA sequencing. *European Heart Journal*, *41*(9), 1024–1036. https://doi.org/10.1093/eurheartj/ehz351

McKenna, A., Findlay, G. M., Gagnon, J. A., Horwitz, M. S., Schier, A. F., & Shendure, J. (2016). Whole-organism lineage tracing by combinatorial and cumulative genome editing. *Science*, *353*(6298). https://doi.org/10.1126/science.aaf7907

McLarren, K. W., Litsiou, A., & Streit, A. (2003). DLX5 positions the neural crest and preplacode region at the border of the neural plate. *Developmental Biology*, *259*(1), 34–47. https://doi.org/10.1016/s0012-1606(03)00177-5

Melsted, P., Booeshaghi, A. S., Liu, L., Gao, F., Lu, L., Min, K. H. J., da Veiga Beltrame, E., Hjörleifsson, K. E., Gehring, J., & Pachter, L. (2021). Modular, efficient and constant-memory single-cell RNA-seq preprocessing. *Nature Biotechnology*, *39*(7), 813–818. https://doi.org/10.1038/s41587-021-00870-2

Menon, M., Mohammadi, S., Davila-Velderrain, J., Goods, B. A., Cadwell, T. D., Xing, Y., Stemmer-Rachamimov, A., Shalek, A. K., Love, J. C., Kellis, M., & Hafler, B. P. (2019). Single-cell transcriptomic atlas of the human retina identifies cell types associated with age-related macular degeneration. *Nature Communications*, *10*(1), 4902. https://doi.org/10.1038/s41467-019-12780-8

Metzker, M. L. (2010). Sequencing technologies — the next generation. *Nature Reviews. Genetics*, *11*(1), 31–46. https://doi.org/10.1038/nrg2626

Michelet, F., Balasankar, A., Teo, N., Stanton, L. W., & Singhal, S. (2020). Rapid generation of purified human RPE from pluripotent stem cells using 2D cultures and lipoprotein uptake-based sorting. *Stem Cell Research & Therapy*, *11*(1), 47. https://doi.org/10.1186/s13287-020-1568-3

Miroshnikova, Y. A., Shahbazi, M. N., Negrete, J., Jr, Chalut, K. J., & Smith, A. (2023). Cell state transitions: catch them if you can. *Development (Cambridge, England)*, *150*(6). https://doi.org/10.1242/dev.201139

Mizukoshi, C., Kojima, Y., Nomura, S., Hayashi, S., Abe, K., & Shimamura, T. (2023). A deep generative model for estimating single-cell RNA splicing and degradation rates. In *bioRxiv*. https://doi.org/10.1101/2023.11.25.568659

Mocellin, S., & Provenzano, M. (2004). RNA interference: learning gene knock-down from cell physiology. *Journal of Translational Medicine*, *2*(1), 39. https://doi.org/10.1186/1479-5876-2-39

Monod, J., Pappenheimer, A. M., Jr, & Cohen-Bazire, G. (1952). La cinétique de la biosynthèse de la β-galactosidase chez E. coli considérée comme fonction de la croissance. *Biochimica et biophysica acta*, *9*, 648–660. https://doi.org/10.1016/0006-3002(52)90227-8

Moon, K. R., Stanley, J. S., III, Burkhardt, D., van Dijk, D., Wolf, G., & Krishnaswamy, S. (2018). Manifold learning-based methods for analyzing single-cell RNA-sequencing data. *Current Opinion in Systems Biology*, *7*, 36–46. https://doi.org/10.1016/j.coisb.2017.12.008

Morris, S. A. (2019). The evolving concept of cell identity in the single cell era. *Development (Cambridge, England)*, *146*(12), dev169748. https://doi.org/10.1242/dev.169748

Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L., & Wold, B. (2008). Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nature Methods*, *5*(7), 621–628. https://doi.org/10.1038/nmeth.1226

Mulas, C., Chaigne, A., Smith, A., & Chalut, K. J. (2021). Cell state transitions: definitions and challenges. *Development (Cambridge, England)*, *148*(20). https://doi.org/10.1242/dev.199950

Müthing, J., Schweppe, C. H., Karch, H., & Friedrich, A. W. (2009). Shiga toxins, glycosphingolipid diversity, and endothelial cell injury. *Thrombosis and Haemostasis*, *101*(2), 252–264. https://www.ncbi.nlm.nih.gov/pubmed/19190807

Nagalakshmi, U., Wang, Z., Waern, K., Shou, C., Raha, D., Gerstein, M., & Snyder, M. (2008). The transcriptional landscape of the yeast genome defined by RNA sequencing. *Science (New York, N.Y.)*, *320*(5881), 1344–1349. https://doi.org/10.1126/science.1158441

Nagiel, A., Lanza, R., & Schwartz, S. D. (2015). Transplantation of Human Embryonic Stem Cell-Derived Retinal Pigment Epithelium for the Treatment of Macular Degeneration. In E. P. Rakoczy (Ed.), *Gene- and Cell-Based Treatment Strategies for the Eye* (pp. 77–86). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-662-45188-5_7

Nasevicius, A., & Ekker, S. C. (2000). Effective targeted gene 'knockdown' in zebrafish. *Nature Genetics*, *26*(2), 216–220. https://doi.org/10.1038/79951

Niehaus, M., Soltwisch, J., Belov, M. E., & Dreisewerd, K. (2019). Transmission-mode MALDI-2 mass spectrometry imaging of cells and tissues at subcellular resolution. *Nature Methods*, *16*(9), 925–931. https://doi.org/10.1038/s41592-019-0536-2

Niemitz, E. (2007). The microarray revolution. *Nature Reviews. Genetics*, *8*(S1), S15–S15. https://doi.org/10.1038/nrg2259

Nishiguchi, S., Wood, H., & Kondoh, H. (1998). Sox1 directly regulates the γ-crystallin genes and is essential for lens development in mice. *Genes*. http://genesdev.cshlp.org/content/12/6/776.short

Norris, J. L., & Caprioli, R. M. (2013). Analysis of tissue specimens by matrix-assisted laser desorption/ionization imaging mass spectrometry in biological and clinical research. *Chemical Reviews*, *113*(4), 2309–2342. https://doi.org/10.1021/cr3004295

Nowakowski, R. S., Lewin, S. B., & Miller, M. W. (1989). Bromodeoxyuridine immunohistochemical determination of the lengths of the cell cycle and the DNA-synthetic phase for an anatomically defined population. *Journal of Neurocytology*, *18*(3), 311–318. https://doi.org/10.1007/BF01190834

Nüsslein-Volhard, C., & Wieschaus, E. (1980). Mutations affecting segment number and polarity in Drosophila. *Nature*, *287*(5785), 795–801. https://doi.org/10.1038/287795a0

Ohnuma, S.-I., & Harris, W. A. (2003). Neurogenesis and the cell cycle. *Neuron*, *40*(2), 199–208. https://doi.org/10.1016/s0896-6273(03)00632-9

Oki, T., Nishimura, K., Kitaura, J., Togami, K., Maehara, A., Izawa, K., Sakaue-Sawano, A., Niida, A., Miyano, S., Aburatani, H., Kiyonari, H., Miyawaki, A., & Kitamura, T. (2014). A novel cell-cycle-indicator, mVenus-p27K–, identifies quiescent cells and visualizes G0–G1 transition. *Scientific Reports*, *4*(1), 1–10. https://doi.org/10.1038/srep04012

Ontology Consortium, G., Aleksander, S. A., Balhoff, J., Carbon, S., Cherry, J. M., & Drabkin, H. J. (2023). The Gene Ontology knowledgebase in 2023. *Genetics*, *224*.

Pachitariu, M., & Stringer, C. (2022). Cellpose 2.0: how to train your own model. *Nature Methods*, *19*(12), 1634–1641. https://doi.org/10.1038/s41592-022-01663-4

Pan, G., & Thomson, J. A. (2007). Nanog and transcriptional networks in embryonic stem cell pluripotency. *Cell Research*, *17*(1), 42–49. https://doi.org/10.1038/sj.cr.7310125

Peidli, S., Green, T. D., Shen, C., Gross, T., Min, J., Garda, S., Yuan, B., Schumacher, L. J., Taylor-King, J. P., Marks, D. S., Luna, A., Blüthgen, N., & Sander, C. (2023). scPerturb: Harmonized Single-Cell Perturbation Data. In *bioRxiv* (p. 2022.08.20.504663). https://doi.org/10.1101/2022.08.20.504663

Pera, M. F., & Rossant, J. (2021). The exploration of pluripotency space: Charting cell state transitions in peri-implantation development. *Cell Stem Cell*, *28*(11), 1896–1906. https://doi.org/10.1016/j.stem.2021.10.001

Petrosius, V., & Schoof, E. M. (2023). Recent advances in the field of single-cell proteomics. *Translational Oncology*, *27*(101556), 101556. https://doi.org/10.1016/j.tranon.2022.101556

Petrus-Reurer, S., Bartuma, H., Aronsson, M., Westman, S., Lanner, F., André, H., & Kvanta, A. (2017). Integration of Subretinal Suspension Transplants of Human Embryonic Stem Cell-Derived Retinal Pigment Epithelial Cells in a Large-Eyed Model of Geographic Atrophy. In *Investigative Opthalmology & Visual Science* (Vol. 58, Issue 2, p. 1314). https://doi.org/10.1167/iovs.16-20738

Petrus-Reurer, S., Bartuma, H., Aronsson, M., Westman, S., Lanner, F., & Kvanta, A. (2018). Subretinal Transplantation of Human Embryonic Stem Cell Derived-retinal Pigment Epithelial Cells into a Large-eyed Model of Geographic Atrophy. *Journal of Visualized Experiments: JoVE*, *131*. https://doi.org/10.3791/56702

Petrus-Reurer, S., Lederer, A. R., Baqué-Vidal, L., Douagi, I., Pannagel, B., Khven, I., Aronsson, M., Bartuma, H., Wagner, M., Wrona, A., Efstathopoulos, P., Jaberi, E., Willenbrock, H., Shimizu, Y., Villaescusa, J. C., André, H., Sundström, E., Bhaduri, A., Kriegstein, A., … Lanner, F. (2022). Molecular profiling of stem cell-derived retinal pigment epithelial cell differentiation established for clinical translation. *Stem Cell Reports*, *17*(6), 1458–1475. https://doi.org/10.1016/j.stemcr.2022.05.005

Philippeos, C., Telerman, S. B., Oulès, B., Pisco, A. O., Shaw, T. J., Elgueta, R., Lombardi, G., Driskell, R. R., Soldin, M., Lynch, M. D., & Watt, F. M. (2018). Spatial and single-cell transcriptional profiling identifies functionally distinct human dermal fibroblast subpopulations. *The Journal of Investigative Dermatology*, *138*(4), 811–825. https://doi.org/10.1016/j.jid.2018.01.016

Piao, J., Zabierowski, S., Dubose, B. N., Hill, E. J., Navare, M., Claros, N., Rosen, S., Ramnarine, K., Horn, C., Fredrickson, C., Wong, K., Safford, B., Kriks, S., El Maarouf, A., Rutishauser, U., Henchcliffe, C., Wang, Y., Riviere, I., Mann, S., … Tabar, V. (2021). Preclinical Efficacy and Safety of a Human Embryonic Stem Cell-Derived Midbrain Dopamine Progenitor Product, MSK-DA01. *Cell Stem Cell*, *28*(2), 217-229.e7. https://doi.org/10.1016/j.stem.2021.01.004

Piwecka, M., Rajewsky, N., & Rybak-Wolf, A. (2023). Single-cell and spatial transcriptomics: deciphering brain complexity in health and disease. *Nature Reviews. Neurology*, *19*(6), 346–362. https://doi.org/10.1038/s41582-023-00809-y

Plaza Reyes, A., Petrus-Reurer, S., Padrell Sanchez, S., Kumar, P., Douagi, I., Bartuma, H., Aronsson, M., Westman, S., Lardner, E., Andre, H., Falk, A., Nandrot, E. F., Kvanta, A., & Lanner, F. (2020). Identification of cell surface markers and establishment of monolayer differentiation to retinal pigment epithelial cells. *Nature Communications*, *11*(1), 1609. https://doi.org/10.1038/s41467-020-15326-5

Plaza Reyes, Alvaro, Petrus-Reurer, S., Antonsson, L., Stenfelt, S., Bartuma, H., Panula, S., Mader, T., Douagi, I., André, H., Hovatta, O., Lanner, F., & Kvanta, A. (2016). Xeno-Free and Defined Human Embryonic Stem Cell-Derived Retinal Pigment Epithelial Cells Functionally Integrate in a Large-Eyed Preclinical Model. *Stem Cell Reports*, *6*(1), 9–17. https://doi.org/10.1016/j.stemcr.2015.11.008

Pope, S. D., & Medzhitov, R. (2018). Emerging principles of gene expression programs and their regulation. *Molecular Cell*, *71*(3), 389–397. https://doi.org/10.1016/j.molcel.2018.07.017

Pozojevic, J., & Spielmann, M. (2023). Single-cell sequencing in neurodegenerative disorders. *Molecular Diagnosis & Therapy*, *27*(5), 553–561. https://doi.org/10.1007/s40291-023-00668-9

Qiao, C., & Huang, Y. (2021). Representation learning of RNA velocity reveals robust cell transitions. *Proceedings of the National Academy of Sciences of the United States of America*, *118*(49). https://doi.org/10.1073/pnas.2105859118

Qin, Q., Bingham, E., La Manno, G., Langenau, D. M., & Pinello, L. (2022). Pyro-Velocity: Probabilistic RNA Velocity inference from single-cell data. In *bioRxiv* (p. 2022.09.12.507691). https://doi.org/10.1101/2022.09.12.507691

Qiu, P. (2020). Embracing the dropouts in single-cell RNA-seq analysis. *Nature Communications*, *11*(1), 1–9. https://doi.org/10.1038/s41467-020-14976-9

Qiu, X., Zhang, Y., Martin-Rufino, J. D., Weng, C., Hosseinzadeh, S., Yang, D., Pogson, A. N., Hein, M. Y., Hoi Joseph Min, K., Wang, L., Grody, E. I., Shurtleff, M. J., Yuan, R., Xu, S., Ma, Y., Replogle, J. M., Lander, E. S., Darmanis, S., Bahar, I., … Weissman, J. S. (2022). Mapping transcriptomic vector fields of single cells. *Cell*, *185*(4), 690-711.e45. https://doi.org/10.1016/j.cell.2021.12.045

Qu, X.-B., Pan, J., Zhang, C., & Huang, S.-Y. (2008). Sox17 facilitates the differentiation of mouse embryonic stem cells into primitive and definitive endoderm in vitro. *Development, Growth & Differentiation*, *50*(7), 585–593. https://doi.org/10.1111/j.1440-169x.2008.01056.x

Quadrato, G., Nguyen, T., Macosko, E. Z., Sherwood, J. L., Min Yang, S., Berger, D. R., Maria, N., Scholvin, J., Goldman, M., Kinney, J. P., Boyden, E. S., Lichtman, J. W., Williams, Z. M., McCarroll, S. A., & Arlotta, P. (2017). Cell diversity and network dynamics in photosensitive human brain organoids. *Nature*, *545*(7652), 48–53. https://doi.org/10.1038/nature22047

Quake, S. R., & Sapiens Consortium, T. (2021). The Tabula Sapiens: a single cell transcriptomic atlas of multiple organs from individual human donors. *Biorxiv*. https://www.biorxiv.org/content/10.1101/2021.07.19.452956.abstract

Quake, Stephen R. (2022). A decade of molecular cell atlases. *Trends in Genetics: TIG*, *38*(8), 805–810. https://doi.org/10.1016/j.tig.2022.01.004

Raj, B., Wagner, D. E., McKenna, A., Pandey, S., Klein, A. M., Shendure, J., Gagnon, J. A., & Schier, A. F. (2018). Simultaneous single-cell profiling of lineages and cell types in the vertebrate brain. *Nature Biotechnology*, *36*(5), 442–450. https://doi.org/10.1038/nbt.4103

Rand, D. A., Raju, A., Sáez, M., Corson, F., & Siggia, E. D. (2021). Geometry of gene regulatory dynamics. *Proceedings of the National Academy of Sciences of the United States of America*, *118*(38), e2109729118. https://doi.org/10.1073/pnas.2109729118

Ranek, J. S., Stanley, N., & Purvis, J. E. (2022). Integrating temporal single-cell gene expression modalities for trajectory inference and disease prediction. *Genome Biology*, *23*(1), 186. https://doi.org/10.1186/s13059-022-02749-0

Rappez, L., Stadler, M., Triana, S., Gathungu, R. M., Ovchinnikova, K., Phapale, P., Heikenwalder, M., & Alexandrov, T. (2021). SpaceM reveals metabolic states of single cells. *Nature Methods*, *18*(7), 799–805. https://doi.org/10.1038/s41592-021-01198-0

Replogle, J. M., Saunders, R. A., Pogson, A. N., Hussmann, J. A., Lenail, A., Guna, A., Mascibroda, L., Wagner, E. J., Adelman, K., Lithwick-Yanai, G., Iremadze, N., Oberstrass, F., Lipson, D., Bonnar, J. L., Jost, M., Norman, T. M., & Weissman, J. S. (2022). Mapping information-rich genotype-phenotype landscapes with genome-scale Perturb-seq. *Cell*, *185*(14), 2559-2575.e28. https://doi.org/10.1016/j.cell.2022.05.013

Reynolds, A. (2007). The cell's journey: from metaphorical to literal factory. *Endeavour*, *31*(2), 65–70. https://doi.org/10.1016/j.endeavour.2007.05.005

Rheaume, B. A., Jereen, A., Bolisetty, M., Sajid, M. S., Yang, Y., Renna, K., Sun, L., Robson, P., & Trakhtenberg, E. F. (2018). Single cell transcriptome profiling of retinal ganglion cells identifies cellular subtypes. *Nature Communications*, *9*(1), 2759. https://doi.org/10.1038/s41467-018-05134-3

Riba, A., Oravecz, A., Durik, M., Jiménez, S., Alunni, V., Cerciat, M., Jung, M., Keime, C., Keyes, W. M., & Molina, N. (2022). Cell cycle gene regulation dynamics revealed by RNA velocity and deep-learning. *Nature Communications*, *13*(1), 2865. https://doi.org/10.1038/s41467-022-30545-8

Rizzolo, L. J., Nasonkin, I. O., & Adelman, R. A. (2022). Retinal cell transplantation, biomaterials, and in vitro models for developing next-generation therapies of age-related macular degeneration. *Stem Cells Translational Medicine*, *11*(3), 269–281. https://doi.org/10.1093/stcltm/szac001

Rognoni, E., Pisco, A. O., Hiratsuka, T., Sipilä, K. H., Belmonte, J. M., Mobasseri, S. A., Philippeos, C., Dilão, R., & Watt, F. M. (2018). Fibroblast state switching orchestrates dermal maturation and wound healing. *Molecular Systems Biology*, *14*(8), e8174. https://doi.org/10.15252/msb.20178174

Rood, J. E., Maartens, A., Hupalowska, A., Teichmann, S. A., & Regev, A. (2022). Impact of the Human Cell Atlas on medicine. *Nature Medicine*, *28*(12), 2486–2496. https://doi.org/10.1038/s41591-022-02104-7

Ross-Macdonald, P., Coelho, P. S., Roemer, T., Agarwal, S., Kumar, A., Jansen, R., Cheung, K. H., Sheehan, A., Symoniatis, D., Umansky, L., Heidtman, M., Nelson, F. K., Iwasaki, H., Hager, K., Gerstein, M., Miller, P., Roeder, G. S., & Snyder, M. (1999). Large-scale analysis of the yeast genome by transposon tagging and gene disruption. *Nature*, *402*(6760), 413–418. https://doi.org/10.1038/46558

Rubner, R., CellSight Ocular Stem Cell and Regeneration Research Program, Department of Ophthalmology, Sue Anschutz-Rodgers Eye Center, University of Colorado School of Medicine, Aurora, CO 80045, USA, Valeria Canto-Soler, M., & CellSight Ocular Stem Cell and Regeneration Research Program, Department of Ophthalmology, Sue Anschutz-Rodgers Eye Center, University of Colorado School of Medicine, Aurora, CO 80045, USA. (2022). Progress of clinical therapies for dry age-related macular degeneration. *International Journal of Ophthalmology*, *15*(1), 157–166. https://doi.org/10.18240/ijo.2022.01.23

Russo, B., Brembilla, N. C., & Chizzolini, C. (2020). Interplay between keratinocytes and fibroblasts: A systematic review providing a new angle for understanding skin fibrotic disorders. *Frontiers in Immunology*, *11*, 648. https://doi.org/10.3389/fimmu.2020.00648

Russo, D., Capolupo, L., Loomba, J. S., Sticco, L., & D'Angelo, G. (2018). Glycosphingolipid metabolism in cell fate specification. *Journal of Cell Science*, *131*(24), jcs219204. https://doi.org/10.1242/jcs.219204

Russo, D., Della Ragione, F., Rizzo, R., Sugiyama, E., Scalabrì, F., Hori, K., Capasso, S., Sticco, L., Fioriniello, S., De Gregorio, R., Granata, I., Guarracino, M. R., Maglione, V., Johannes, L., Bellenchi, G. C., Hoshino, M., Setou, M., D'Esposito, M., Luini, A., & D'Angelo, G. (2018). Glycosphingolipid metabolic reprogramming drives neural differentiation. *The EMBO Journal*, *37*(7), e97674. https://doi.org/10.15252/embj.201797674

Saelens, W., Cannoodt, R., Todorov, H., & Saeys, Y. (2019). A comparison of single-cell trajectory inference methods. *Nature Biotechnology*, *37*(5), 547–554. https://doi.org/10.1038/s41587-019-0071-9

Sáez, M., Briscoe, J., & Rand, D. A. (2022). Dynamical landscapes of cell fate decisions. *Interface Focus*, *12*(4). https://doi.org/10.1098/rsfs.2022.0002

Salero, E., Blenkinsop, T. A., Corneo, B., Harris, A., Rabin, D., Stern, J. H., & Temple, S. (2012). Adult human RPE can be activated into a multipotent stem cell that produces mesenchymal derivatives. *Cell Stem Cell*, *10*(1), 88–95. https://doi.org/10.1016/j.stem.2011.11.018

Santillán, M., & Mackey, M. C. (2008). Quantitative approaches to the study of bistability in the *lac* operon of *Escherichia coli*. *Journal of the Royal Society, Interface*, *5*(suppl_1), S29. https://doi.org/10.1098/rsif.2008.0086.focus

Santos, A., Wernersson, R., & Jensen, L. J. (2015). Cyclebase 3.0: a multi-organism database on cell-cycle regulation and phenotypes. *Nucleic Acids Research*, *43*(Database issue), D1140-4. https://doi.org/10.1093/nar/gku1092

Sarkar, A., Marchetto, M. C., & Gage, F. H. (2020). Chapter 9 - Neural induction of embryonic stem/induced pluripotent stem cells. In J. Rubenstein, P. Rakic, B. Chen, & K. Y. Kwan (Eds.), *Patterning and Cell Type Specification in the Developing CNS and PNS (Second Edition)* (pp. 185–203). Academic Press. https://doi.org/10.1016/B978-0-12-814405-3.00009-6

Satija, R., Farrell, J. A., Gennert, D., Schier, A. F., & Regev, A. (2015). Spatial reconstruction of single-cell gene expression data. *Nature Biotechnology*, *33*(5), 495–502. https://doi.org/10.1038/nbt.3192

Schena, M., Shalon, D., Davis, R. W., & Brown, P. O. (1995). Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science (New York, N.Y.)*, *270*(5235), 467–470. https://doi.org/10.1126/science.270.5235.467

Schmidt, F., Cherepkova, M. Y., & Platt, R. J. (2018). Transcriptional recording by CRISPR spacer acquisition from RNA. *Nature*, *562*(7727), 380–385. https://doi.org/10.1038/s41586-018-0569-1

Schmidt, F., Zimmermann, J., Tanna, T., Farouni, R., Conway, T., Macpherson, A. J., & Platt, R. J. (2022). Noninvasive assessment of gut function using transcriptional recording sentinel cells. *Science (New York, N.Y.)*, *376*(6594). https://doi.org/10.1126/science.abm6038

Schmidt, U., Weigert, M., Broaddus, C., & Myers, G. (2018). Cell Detection with Star-convex Polygons. In *arXiv [cs.CV]*. arXiv. http://arxiv.org/abs/1806.03535

Schmitt, S., Aftab, U., Jiang, C., Redenti, S., Klassen, H., Miljan, E., Sinden, J., & Young, M. (2009). Molecular characterization of human retinal progenitor cells. *Investigative Ophthalmology & Visual Science*, *50*(12), 5901–5908. https://doi.org/10.1167/iovs.08-3067

Schober, Y., Guenther, S., Spengler, B., & Römpp, A. (2012). Single cell matrix-assisted laser desorption/ionization mass spectrometry imaging. *Analytical Chemistry*, *84*(15), 6293–6297. https://doi.org/10.1021/ac301337h

Schwabe, D., Formichetti, S., Junker, J. P., Falcke, M., & Rajewsky, N. (2020). The transcriptome dynamics of single cells during the cell cycle. *Molecular Systems Biology*, *16*(11), e9946. https://doi.org/10.15252/msb.20209946

Seo, S., Chen, L., Liu, W., Zhao, D., Schultz, K. M., Sasman, A., Liu, T., Zhang, H. F., Gage, P. J., & Kume, T. (2017). Foxc1 and Foxc2 in the Neural Crest Are Required for Ocular Anterior Segment Development. *Investigative Ophthalmology & Visual Science*, *58*(3), 1368–1377. https://doi.org/10.1167/iovs.16-21217

Shao, J., Zhou, P.-Y., & Peng, G.-H. (2017). Experimental Study of the Biological Properties of Human Embryonic Stem Cell-Derived Retinal Progenitor Cells. *Scientific Reports*, *7*, 42363. https://doi.org/10.1038/srep42363

Sharma, A., Takata, H., Shibahara, K.-I., Bubulya, A., & Bubulya, P. A. (2010). Son is essential for nuclear speckle organization and cell cycle progression. *Molecular Biology of the Cell*, *21*(4), 650–663. https://doi.org/10.1091/mbc.e09-02-0126

Sharma, R., Khristov, V., Rising, A., Jha, B. S., Dejene, R., Hotaling, N., Li, Y., Stoddard, J., Stankewicz, C., Wan, Q., Zhang, C., Campos, M. M., Miyagishima, K. J., McGaughey, D., Villasmil, R., Mattapallil, M., Stanzel, B., Qian, H., Wong, W., … Bharti, K. (2019). Clinical-grade stem cell-derived retinal pigment epithelium patch rescues retinal degeneration in rodents and pigs. *Science Translational Medicine*, *11*(475). https://doi.org/10.1126/scitranslmed.aat5580

Shekhar, K., Lapan, S. W., Whitney, I. E., Tran, N. M., Macosko, E. Z., Kowalczyk, M., Adiconis, X., Levin, J. Z., Nemesh, J., Goldman, M., McCarroll, S. A., Cepko, C. L., Regev, A., & Sanes, J. R. (2016). Comprehensive Classification of Retinal Bipolar Neurons by Single-Cell Transcriptomics. *Cell*, *166*(5), 1308-1323.e30. https://doi.org/10.1016/j.cell.2016.07.054

Sheng, Y., Barak, B., & Nitzan, M. (2023). Robust reconstruction of single-cell RNA-seq data with iterative gene weight updates. *Bioinformatics* , *39*(39 Suppl 1), i423–i430. https://doi.org/10.1093/bioinformatics/btad253

Shen-Orr, S. S., Milo, R., Mangan, S., & Alon, U. (2002). Network motifs in the transcriptional regulation network of Escherichia coli. *Nature Genetics*, *31*(1), 64–68. https://doi.org/10.1038/ng881

Shi, Y., Inoue, H., Wu, J. C., & Yamanaka, S. (2017). Induced pluripotent stem cell technology: a decade of progress. *Nature Reviews. Drug Discovery*, *16*(2), 115–130. https://doi.org/10.1038/nrd.2016.245

Snijder, B., Sacher, R., Rämö, P., Damm, E.-M., Liberali, P., & Pelkmans, L. (2009). Population context determines cell-to-cell variability in endocytosis and virus infection. *Nature*, *461*(7263), 520–523. https://doi.org/10.1038/nature08282

Sofroniew, N., Lambert, T., Evans, K., Nunez-Iglesias, J., Bokota, G., Peña-Castellanos, G., Winston, P., Yamauchi, K., Bussonnier, M., Pop, D. D., Liu, Z., ACS, Pam, alisterburt, Buckley, G., Sweet, A., Gaifas, L., Rodríguez-Guerra, J., Migas, L., … Har-Gil, H. (2021). *napari/napari: 0.4.12rc2*. Zenodo. https://doi.org/10.5281/ZENODO.3555620

Soldatov, R., Kaucka, M., Kastriti, M. E., & Petersen, J. (2019). *Spatiotemporal structure of cell fate decisions in murine neural crest*. https://science.sciencemag.org/content/364/6444/eaas9536.abstract?casa_token=V14TUFpFE0gAAAAA:8OgNMKz77P_p6b2ZhDTtE6yfld65tHvu1S7Ch7LUQQyXbQNG4ps3p5wDeAlmtAsRZiQvpedVDkIHzp8

Song, W. K., Park, K.-M., Kim, H.-J., Lee, J. H., Choi, J., Chong, S. Y., Shim, S. H., Del Priore, L. V., & Lanza, R. (2015). Treatment of macular degeneration using embryonic stem cell-derived retinal pigment epithelium: preliminary results in Asian patients. *Stem Cell Reports*, *4*(5), 860–872. https://doi.org/10.1016/j.stemcr.2015.04.005

Spanjaard, B., Hu, B., Mitic, N., Olivares-Chauvet, P., Janjuha, S., Ninov, N., & Junker, J. P. (2018). Simultaneous lineage tracing and cell-type identification using CRISPR-Cas9-induced genetic scars. *Nature Biotechnology*, *36*(5), 469–473. https://doi.org/10.1038/nbt.4124

Sparrow, J. R., Hicks, D., & Hamel, C. P. (2010). The retinal pigment epithelium in health and disease. *Current Molecular Medicine*, *10*(9), 802–823. https://doi.org/10.2174/156652410793937813

Spemann & Mangold. (1923). *Über Induktion von Embryonalanlagen durch Implantation artfremder Organisatoren*. Archiv für Mikroskopische Anatomie und Entwicklungsmechanik.

Sridhar, A., Hoshino, A., Finkbeiner, C. R., Chitsazan, A., Dai, L., Haugan, A. K., Eschenbacher, K. M., Jackson, D. L., Trapnell, C., Bermingham-McDonogh, O., Glass, I., & Reh, T. A. (2020). Single-Cell Transcriptomic Comparison of Human Fetal Retina, hPSC-Derived Retinal Organoids, and Long-Term Retinal Cultures. In *Cell Reports* (Vol. 30, Issue 5, pp. 1644-1659.e4). https://doi.org/10.1016/j.celrep.2020.01.007

Stoeckius, M., Hafemeister, C., Stephenson, W., Houck-Loomis, B., Chattopadhyay, P. K., Swerdlow, H., Satija, R., & Smibert, P. (2017). Simultaneous epitope and transcriptome measurement in single cells. *Nature Methods*, *14*(9), 865–868. https://doi.org/10.1038/nmeth.4380

Streit, A. (2007). The preplacodal region: an ectodermal domain with multipotential progenitors that contribute to sense organs and cranial sensory ganglia. *The International Journal of Developmental Biology*, *51*(6–7), 447–461. https://doi.org/10.1387/ijdb.072327as

Stringer, C., Wang, T., Michaelos, M., & Pachitariu, M. (2021). Cellpose: a generalist algorithm for cellular segmentation. *Nature Methods*, *18*(1), 100–106. https://doi.org/10.1038/s41592-020-01018-x

Sugiyama, E., Yao, I., & Setou, M. (2018). Visualization of local phosphatidylcholine synthesis within hippocampal neurons using a compartmentalized culture system and imaging mass spectrometry. *Biochemical and Biophysical Research Communications*, *495*(1), 1048–1054. https://doi.org/10.1016/j.bbrc.2017.11.108

Sunness, J. S. (1999). The natural history of geographic atrophy, the advanced atrophic form of age-related macular degeneration. *Molecular Vision*, *5*, 25. https://www.ncbi.nlm.nih.gov/pubmed/10562649

Sutter, A. P., Höpfner, M., Huether, A., Maaser, K., & Scherübl, H. (2006). Targeting the epidermal growth factor receptor by erlotinib (Tarceva) for the treatment of esophageal cancer. *International Journal of Cancer. Journal International Du Cancer*, *118*(7), 1814–1822. https://doi.org/10.1002/ijc.21512

Svensson, V., da Veiga Beltrame, E., & Pachter, L. (2020). A curated database reveals trends in single-cell transcriptomics. *Database: The Journal of Biological Databases and Curation*, *2020*. https://doi.org/10.1093/database/baaa073

Svensson, V., & Pachter, L. (2018). RNA Velocity: Molecular Kinetics from Single-Cell RNA-Seq [Review of *RNA Velocity: Molecular Kinetics from Single-Cell RNA-Seq*]. *Molecular Cell*, *72*(1), 7–9. https://doi.org/10.1016/j.molcel.2018.09.026

Tahayato, A., Dollé, P., & Petkovich, M. (2003). Cyp26C1 encodes a novel retinoic acid-metabolizing enzyme expressed in the hindbrain, inner ear, first branchial arch and tooth buds during murine development. *Gene Expression Patterns: GEP*, *3*(4), 449–454. https://doi.org/10.1016/s1567-133x(03)00066-8

Takahashi, K., & Yamanaka, S. (2006). Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors. *Cell*, *126*(4), 663–676. https://doi.org/10.1016/j.cell.2006.07.024

Talamanca, L., Gobet, C., & Naef, F. (2023). Sex-dimorphic and age-dependent organization of 24-hour gene expression rhythms in humans. *Science (New York, N.Y.)*, *379*(6631), 478–483. https://doi.org/10.1126/science.add0846

Talamanca, L., & Naef, F. (2020). How to tell time: advances in decoding circadian phase from omics snapshots. *F1000Research*, *9*, 1150. https://doi.org/10.12688/f1000research.26759.1

Tang, F., Barbacioru, C., Wang, Y., Nordman, E., Lee, C., Xu, N., Wang, X., Bodeau, J., Tuch, B. B., Siddiqui, A., Lao, K., & Surani, M. A. (2009). mRNA-Seq whole-transcriptome analysis of a single cell. *Nature Methods*, *6*(5), 377–382. https://doi.org/10.1038/nmeth.1315

Tang, W., & Liu, D. R. (2018). Rewritable multi-event analog recording in bacterial and mammalian cells. *Science*, *360*(6385), 8992. https://doi.org/10.1126/science.aap8992

Tedesco, M., Giannese, F., Lazarević, D., Giansanti, V., Rosano, D., Monzani, S., Catalano, I., Grassi, E., Zanella, E. R., Botrugno, O. A., Morelli, L., Panina Bordignon, P., Caravagna, G., Bertotti, A., Martino, G., Aldrighetti, L., Pasqualato, S., Trusolino, L., Cittaro, D., & Tonon, G. (2022). Chromatin Velocity reveals epigenetic dynamics by single-cell profiling of heterochromatin and euchromatin. *Nature Biotechnology*, *40*(2), 235–244. https://doi.org/10.1038/s41587-021-01031-1

Theilgaard-Mönch, K., Pundhir, S., Reckzeh, K., Su, J., Tapia, M., Furtwängler, B., Jendholm, J., Jakobsen, J. S., Hasemann, M. S., Knudsen, K. J., Cowland, J. B., Fossum, A., Schoof, E., Schuster, M. B., & Porse, B. T. (2022). Transcription factor-driven coordination of cell cycle exit and lineage-specification in vivo during granulocytic differentiation. *Nature Communications*, *13*(1), 1–17. https://doi.org/10.1038/s41467-022-31332-1

Thiele, C., Wunderling, K., & Leyendecker, P. (2019). Multiplexed and single cell tracing of lipid metabolism. *Nature Methods*, *16*(11), 1123–1130. https://doi.org/10.1038/s41592-019-0593-6

Thomas, M. B., Chadha, R., Glover, K., Wang, X., Morris, J., Brown, T., Rashid, A., Dancey, J., & Abbruzzese, J. L. (2007). Phase 2 study of erlotinib in patients with unresectable hepatocellular carcinoma. *Cancer*, *110*(5), 1059–1067. https://doi.org/10.1002/cncr.22886

Tian, L., Jabbari, J. S., Thijssen, R., Gouil, Q., Amarasinghe, S. L., Voogd, O., Kariyawasam, H., Du, M. R. M., Schuster, J., Wang, C., Su, S., Dong, X., Law, C. W., Lucattini, A., Prawer, Y. D. J., Collar-Fernández, C., Chung, J. D., Naim, T., Chan, A., … Ritchie, M. E. (2021). Comprehensive characterization of single-cell full-length isoforms in human and mouse with long-read sequencing. *Genome Biology*, *22*(1). https://doi.org/10.1186/s13059-021-02525-6

Tirosh, I., Izar, B., Prakadan, S. M., Wadsworth, M. H., 2nd, Treacy, D., Trombetta, J. J., Rotem, A., Rodman, C., Lian, C., Murphy, G., Fallahi-Sichani, M., Dutton-Regester, K., Lin, J.-R., Cohen, O., Shah, P., Lu, D., Genshaft, A. S., Hughes, T. K., Ziegler, C. G. K., … Garraway, L. A. (2016). Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-seq. *Science*, *352*(6282), 189–196. https://doi.org/10.1126/science.aad0501

Tran, H. T. N., Ang, K. S., Chevrier, M., Zhang, X., Lee, N. Y. S., Goh, M., & Chen, J. (2020). A benchmark of batch-effect correction methods for single-cell RNA sequencing data. *Genome Biology*, *21*(1). https://doi.org/10.1186/s13059-019-1850-9

Trapnell, C., Cacchiarelli, D., Grimsby, J., Pokharel, P., Li, S., Morse, M., Lennon, N. J., Livak, K. J., Mikkelsen, T. S., & Rinn, J. L. (2014). The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nature Biotechnology*, *32*(4), 381–386. https://doi.org/10.1038/nbt.2859

Tritschler, S., Büttner, M., Fischer, D. S., Lange, M., Bergen, V., Lickert, H., & Theis, F. J. (2019). Concepts and limitations for learning developmental trajectories from single cell genomics. *Development (Cambridge, England)*, *146*(12), dev170506. https://doi.org/10.1242/dev.170506

Tyson, J. J., & Novák, B. (2022). Time-keeping and decision-making in the cell cycle. *Interface Focus*, *12*(4), 20210075. https://doi.org/10.1098/rsfs.2021.0075

Ullrich, R. T., Zander, T., Neumaier, B., Koker, M., Shimamura, T., Waerzeggers, Y., Borgman, C. L., Tawadros, S., Li, H., Sos, M. L., Backes, H., Shapiro, G. I., Wolf, J., Jacobs, A. H., Thomas, R. K., & Winkeler, A. (2008). Early detection of erlotinib treatment response in NSCLC by 3'-deoxy-3'-[F]-fluoro-L-thymidine

([F]FLT) positron emission tomography (PET). *PloS One*, *3*(12), e3908. https://doi.org/10.1371/journal.pone.0003908

Van Regenmortel, M. H. V. (2004). Reductionism and complexity in molecular biology. *EMBO Reports*, *5*(11), 1016–1020. https://doi.org/10.1038/sj.embor.7400284

Vandereyken, K., Sifrim, A., Thienpont, B., & Voet, T. (2023). Methods and applications for single-cell and spatial multi-omics. *Nature Reviews. Genetics*, *24*(8), 494–515. https://doi.org/10.1038/s41576-023-00580-2

Velasco, S., Kedaigle, A. J., Simmons, S. K., Nash, A., Rocha, M., Quadrato, G., Paulsen, B., Nguyen, L., Adiconis, X., Regev, A., Levin, J. Z., & Arlotta, P. (2019). Individual brain organoids reproducibly form cell diversity of the human cerebral cortex. *Nature*, *570*(7762), 523–527. https://doi.org/10.1038/s41586-019-1289-x

Veres, A., Faust, A. L., Bushnell, H. L., Engquist, E. N., Kenty, J. H.-R., Harb, G., Poh, Y.-C., Sintov, E., Gürtler, M., Pagliuca, F. W., Peterson, Q. P., & Melton, D. A. (2019). Charting cellular identity during human in vitro β-cell differentiation. *Nature*, *569*(7756), 368–373. https://doi.org/10.1038/s41586-019-1168-5

Vinsland, E., & Linnarsson, S. (2022). Single-cell RNA-sequencing of mammalian brain development: insights and future directions. *Development (Cambridge, England)*, *149*(10), dev200180. https://doi.org/10.1242/dev.200180

Virchow, R. (1858). *Die Cellularpathologie in ihrer Begründung auf physiologische and pathologische Gewebelehre*. Berliner Pathologischen Institut.

Voigt, A. P., Mulfaul, K., Mullin, N. K., Flamme-Wiese, M. J., Giacalone, J. C., Stone, E. M., Tucker, B. A., Scheetz, T. E., & Mullins, R. F. (2019). Single-cell transcriptomics of the human retinal pigment epithelium and choroid in health and macular degeneration. *Proceedings of the National Academy of Sciences of the United States of America*, *116*(48), 24100–24107. https://doi.org/10.1073/pnas.1914143116

Waddington, C. H. (1957). *The Strategy of the Genes; a Discussion of Some Aspects of Theoretical Biology*. Allen & Unwin.

Wagner, D. E., & Klein, A. M. (2020). Lineage tracing meets single-cell omics: opportunities and challenges. *Nature Reviews. Genetics*, *21*(7), 410–427. https://doi.org/10.1038/s41576-020-0223-2

Wang, Y., Tang, Z., & Gu, P. (2020). Stem/progenitor cell-based transplantation for retinal degeneration: a review of clinical trials. *Cell Death & Disease*, *11*(9), 793. https://doi.org/10.1038/s41419-020-02955-3

Weigert, M., Schmidt, U., Haase, R., Sugawara, K., & Myers, G. (2020, March). Star-convex polyhedra for 3D object detection and segmentation in microscopy. *2020 IEEE Winter Conference on Applications of Computer Vision (WACV)*. 2020 IEEE Winter Conference on Applications of Computer Vision (WACV), Snowmass Village, CO, USA. https://doi.org/10.1109/wacv45572.2020.9093435

Weinreb, C., Rodriguez-Fraticelli, A., Camargo, F. D., & Klein, A. M. (2020). Lineage tracing on transcriptional landscapes links state to fate during differentiation. *Science (New York, N.Y.)*, *367*(6479), eaaw3381. https://doi.org/10.1126/science.aaw3381

Weinreb, C., Wolock, S., Tusi, B. K., Socolovsky, M., & Klein, A. M. (2018). Fundamental limits on dynamic inference from single-cell snapshots. *Proceedings of the National Academy of Sciences of the United States of America*, *115*(10), E2467–E2476. https://doi.org/10.1073/pnas.1714723115

Weng, G., Kim, J., & Won, K. J. (2021). VeTra: a tool for trajectory inference based on RNA velocity. *Bioinformatics* , *37*(20), 3509–3513. https://doi.org/10.1093/bioinformatics/btab364

Werner, S., Krieg, T., & Smola, H. (2007). Keratinocyte-fibroblast interactions in wound healing. *The Journal of Investigative Dermatology*, *127*(5), 998–1008. https://doi.org/10.1038/sj.jid.5700786

Wiman, K. G., & Zhivotovsky, B. (2017). Understanding cell cycle and cell death regulation provides novel weapons against human diseases. *Journal of Internal Medicine*, *281*(5), 483–495. https://doi.org/10.1111/joim.12609

Wold, S., Esbensen, K., & Geladi, P. (1987). Principal component analysis. *Chemometrics and Intelligent Laboratory Systems: An International Journal Sponsored by the Chemometrics Society*, *2*(1–3), 37–52. https://doi.org/10.1016/0169-7439(87)80084-9

Wolf, F. A., Angerer, P., & Theis, F. J. (2018). SCANPY: large-scale single-cell gene expression data analysis. *Genome Biology*, *19*(1), 15. https://doi.org/10.1186/s13059-017-1382-0

Wolf, F. A., Hamey, F. K., Plass, M., Solana, J., Dahlin, J. S., Göttgens, B., Rajewsky, N., Simon, L., & Theis, F. J. (2019). PAGA: graph abstraction reconciles clustering with trajectory inference through a topology preserving map of single cells. *Genome Biology*, *20*(1), 59. https://doi.org/10.1186/s13059-019-1663-x

Wolfien, M., Galow, A.-M., Müller, P., Bartsch, M., Brunner, R. M., Goldammer, T., Wolkenhauer, O., Hoeflich, A., & David, R. (2020). Single-nucleus sequencing of an entire mammalian heart: Cell type composition and velocity. *Cells (Basel, Switzerland)*, *9*(2), 318. https://doi.org/10.3390/cells9020318

Wong, J. H. C., Ma, J. Y. W., Jobling, A. I., Brandli, A., Greferath, U., Fletcher, E. L., & Vessey, K. A. (2022). Exploring the pathogenesis of age-related macular degeneration: A review of the interplay between retinal pigment epithelium dysfunction and the innate immune system. *Frontiers in Neuroscience*, *16*. https://doi.org/10.3389/fnins.2022.1009599

Yamada, R., Mizutani-Koseki, Y., Hasegawa, T., Osumi, N., Koseki, H., & Takahashi, N. (2003). Cell-autonomous involvement of Mab21l1 is essential for lens placode development. *Development* , *130*(9), 1759–1770. https://doi.org/10.1242/dev.00399

Yang, S., Zhou, J., & Li, D. (2021). Functions and diseases of the retinal pigment epithelium. *Frontiers in Pharmacology*, *12*. https://doi.org/10.3389/fphar.2021.727870

Yao, M., Ren, T., Pan, Y., Xue, X., Li, R., Zhang, L., Li, Y., & Huang, K. (2022). A new generation of Lineage Tracing dynamically records cell fate choices. *International Journal of Molecular Sciences*, *23*(9), 5021. https://doi.org/10.3390/ijms23095021

Ye, F., Huang, W., & Guo, G. (2017). Studying hematopoiesis using single-cell technologies. *Journal of Hematology & Oncology*, *10*(1). https://doi.org/10.1186/s13045-017-0401-7

Yoon, J. H., Seo, Y., Jo, Y. S., Lee, S., Cho, E., Cazenave-Gassiot, A., Shin, Y.-S., Moon, M. H., An, H. J., Wenk, M. R., & Suh, P.-G. (2022). Brain lipidomics: From functional landscape to clinical significance. *Science Advances*, *8*(37). https://doi.org/10.1126/sciadv.adc9317

Yu, T., & Scolnick, J. (2022). Complex biological questions being addressed using single cell sequencing technologies. *SLAS Technology*, *27*(2), 143–149. https://doi.org/10.1016/j.slast.2021.10.013

Yun, S., Saijoh, Y., Hirokawa, K. E., Kopinke, D., Murtaugh, L. C., Monuki, E. S., & Levine, E. M. (2009). Lhx2 links the intrinsic and extrinsic factors that control optic cup formation. *Development* , *136*(23), 3895–3906. https://doi.org/10.1242/dev.041202

Zarbin, M., Sugino, I., & Townes-Anderson, E. (2019). Concise review: update on retinal pigment epithelium transplantation for age-related macular degeneration. *Stem Cells Translational Medicine*, *8*(5), 466–477. https://stemcellsjournals.onlinelibrary.wiley.com/doi/abs/10.1002/sctm.18-0282@10.1002/(ISSN)2157-6580.wscr-2019

Zavalin, A., Todd, E. M., Rawhouser, P. D., Yang, J., Norris, J. L., & Caprioli, R. M. (2012). Direct imaging of single cells and tissue at sub-cellular spatial resolution using transmission geometry MALDI MS. *Journal of Mass Spectrometry*, *47*(11), 1473–1481. https://doi.org/10.1002/jms.3108

Zeisel, A., Hochgerner, H., Lönnerberg, P., Johnsson, A., Memic, F., van der Zwan, J., Häring, M., Braun, E., Borm, L. E., La Manno, G., Codeluppi, S., Furlan, A., Lee, K., Skene, N., Harris, K. D., Hjerling-Leffler, J., Arenas, E., Ernfors, P., Marklund, U., & Linnarsson, S. (2018). Molecular Architecture of the Mouse Nervous System. *Cell*, *174*(4), 999-1014.e22. https://doi.org/10.1016/j.cell.2018.06.021

Zeisel, A., Köstler, W. J., Molotski, N., Tsai, J. M., Krauthgamer, R., Jacob-Hirsch, J., Rechavi, G., Soen, Y., Jung, S., Yarden, Y., & Others. (2011). Coupled pre-mRNA and mRNA dynamics unveil operational strategies underlying transcriptional responses to stimuli. *Molecular Systems Biology*, *7*(1). https://www.embopress.org/doi/abs/10.1038/msb.2011.62

Zenobi, R. (2013). Single-cell metabolomics: analytical and biological perspectives. *Science (New York, N.Y.)*, *342*(6163), 1243259. https://doi.org/10.1126/science.1243259

Zheng, S. C., Stein-O'Brien, G., Augustin, J. J., Slosberg, J., Carosso, G. A., Winer, B., Shin, G., Bjornsson, H. T., Goff, L. A., & Hansen, K. D. (2022). Universal prediction of cell-cycle position using transfer learning. *Genome Biology*, *23*(1), 41. https://doi.org/10.1186/s13059-021-02581-y

Zheng, S. C., Stein-O'Brien, G., Boukas, L., Goff, L. A., & Hansen, K. D. (2023). Pumping the brakes on RNA velocity by understanding and interpreting RNA velocity estimates. *Genome Biology*, *24*(1), 246. https://doi.org/10.1186/s13059-023-03065-x

Zhou, J., Wang, C., Wang, Z., Dampier, W., Wu, K., Casimiro, M. C., Chepelev, I., Popov, V. M., Quong, A., Tozeren, A., Zhao, K., Lisanti, M. P., & Pestell, R. G. (2010). Attenuation of Forkhead signaling by the retinal determination factor DACH1. *Proceedings of the National Academy of Sciences of the United States of America*, *107*(15), 6864–6869. https://doi.org/10.1073/pnas.1002746107

Zhu, C., Yu, M., Huang, H., Juric, I., Abnousi, A., Hu, R., Lucero, J., Behrens, M. M., Hu, M., & Ren, B. (2019). An ultra high-throughput method for single-cell joint analysis of open chromatin and transcriptome. *Nature Structural & Molecular Biology*, *26*(11), 1063–1070. https://doi.org/10.1038/s41594-019-0323-x

Zimmer, A. M., Pan, Y. K., Chandrapalan, T., Kwong, R. W. M., & Perry, S. F. (2019). Loss-of-function approaches in comparative physiology: is there a future for knockdown experiments in the era of genome editing? *The Journal of Experimental Biology*, *222*(7), jeb175737. https://doi.org/10.1242/jeb.175737

Alex Russell Lederer

Graduate Student and Systems Biologist at EPFL
E-mail: alex.lederer@epfl.ch
linkedin | twitter | scholar

About Me

I am a final year doctoral student in the Computational and Quantitative Biology Program at EPFL. I am fascinated by the rapid growth of systems biology and bioinformatics. My long-term goal is to contribute to the shift in biology from a descriptive to predictive science. I am eager for opportunities to explore new ideas and strengthen my skills across Switzerland and Europe.

Education

PhD Program in Computational and Quantitative Biology 07/2019 – Present
Swiss Federal Institute of Technology (EPFL), Lausanne Switzerland

BS Honors in Molecular Biology and Computer Science, Chemistry minor 08/2013 – 04/2017
University of Pittsburgh Honors College, PA, United States (GPA: 3.68/4.00)

NYS Regents Honors Diploma, AP Scholar
Smithtown High School East, NY, United States (GPA: 4.0/4.0) 09/2009 – 06/2013

Skills

Computational: Python programming, data analysis, bioinformatics, Pyro, variational inference, machine learning, algorithm implementation, next generation sequencing data, software engineering, Git/GitHub version control, Jupyter notebooks, conda
Experimental: single cell technologies (10X Genomics: scRNA-seq and scATAC-seq), molecular biology methods (western and northern blotting, PCR, cloning), yeast genetics, cell culture
Languages: English (native), German (A2)

Work Experience

Doctoral Assistant, Advisor: Dr. Gioele La Manno 07/2019 - Present
Swiss Federal Institute of Technology (EPFL) Lausanne, Switzerland

Project Title: Deciphering cell state transitions in single cells using quantitative temporal models
Applying single-cell tools to explore retinal differentiation in the context of neurodegenerative disease. Developing novel computational methods for RNA velocity of the cell cycle and lipid state transitions using Bayesian inference and Markov chains. Working at the interface of biology and computation.

DAAD Research Fellow, Advisor: Dr. Lars Steinmetz 10/2017 - 07/2018
European Molecular Biology Laboratory Heidelberg, Germany

Project Title: Next-Generation Functional Genomics with Highly Multiplexed Precision Editing
Designed computational pipelines for an improved CRISPR/Cas9 editing system to perturb and probe the genotype-environment-phenotype relationship. Funded by the German Academic Exchange Service.

Undergraduate Researcher, Advisor: Dr. Karen Arndt 01/2014 - 05/2017
University of Pittsburgh Pittsburgh, United States

Project Title: Exploring the Role of the Paf1 Complex in Transcription Regulation
Surveyed the impact of the yeast Paf1 complex on noncoding gene regulation. Defended an honors thesis.

Computational Biology Intern 05/2015 - 08/2015
Asuragen, Inc. Austin, United States

Project Title: Designing a Python Module for Retrieval and Analysis of Multiomics Cancer Atlas Data

Additional Work Experience

Innosuisse Business Concept Training, EPFL Innovation Park 09/2022 - 12/2022
Fast track and hands-on entrepreneurship skills taught by seasoned entrepreneurs: pitch training, market analysis, value proposition, financial planning, presentation strategy.

Freelance Editor, Urban Connections 07/2017 - 07/2019
Editing and transcription of English documents on science and technology for Japanese clients.

Scientific Publications

Statistical inference with a manifold-constrained RNA velocity model uncovers cell cycle speed modulations. **Lederer, A.R.**, Leonardi, M.*, Talamanca, L.*, [*and 5 others*], Naef, F., La Manno, G. **bioRxiv** (2024).

Contributions: **first author**; conceptualized the project; designed and implemented a generative Bayesian model for RNA velocity in Pyro; performed all computational analyses; created figures; wrote and revised manuscript; managed project.

Cryopreservation of differentiated retinal pigment epithelial cells for the treatment of age macular degeneration. Baqué-Vidal, L.*, Main, H.*, Petrus-Reurer, S., **Lederer, A.R.**, [*and 10 others*], Lanner, F. (under review) (2023)

Contributions: performed single-cell RNA sequencing analysis; designed some figure panels.

Molecular profiling of stem cell-derived retinal pigment epithelial cell differentiation established for clinical translation. Petrus-Reurer, S*., **Lederer, A. R.***, Baqué-Vidal, L., [*and 17 others*], La Manno, G., Lanner, F. **Stem Cell Reports** (2022): *10.1016/j.stemcr.2022.05.005* **(featured on cover)**.

Contributions: **co-first author**; conceptualized the project; performed all computational analyses; created all figures; wrote and revised manuscript; managed project.

Sphingolipids Control Dermal Fibroblast Heterogeneity. Capolupo, L., Khven, I., **Lederer, A.R.**, [*and 18 others*], La Manno, G., D'Angelo, G. **Science** (2022): *10.1126/science.abh1623*

Contributions: developed a Markov chain model called CELLMA for predicting single cell lipotype state transitions; performed single cell lipidomics data analysis; revised manuscript.

Molecular architecture of the developing mouse brain. La Manno, G.,* Siletti, K.*, Furlan, A., Gyllborg, D., Vinsland, E., Albiach, A. M., Langseth, C. M., Khven, I., **Lederer, A. R.**, [*and 4 others*], Linnarsson, S. **Nature** (2021): *10.1038/s41586-021-03775-x*

Contributions: assisted with sensory tissue annotations in spatial transcriptomics (HybISS) data.

The emergence and promise of single cell temporal-omics approaches. **Lederer, A. R.**, La Manno, G. **Current Opinion in Biotechnology** (2020): *10.1016/j.copbio.2019.12.005*.

Contributions: conceptualized, wrote, and revised review article.

The Paf1 complex broadly impacts the transcriptome of Saccharomyces cerevisiae. Ellison, M. A., **Lederer, A. R.**, Warner, M. H., [*and 5 others*], Arndt, K. A. **Genetics** (2019): *10.1534/genetics.119.302262*.

Contributions: conducted microarray analysis and northern blots experiments; conceptualized project; revised manuscript.

Multiplexed precision genome editing with trackable genomic barcodes in yeast. Roy, K. R.*, Smith, J. D.*, Vonesch, S. C.*, Lin, G., Szu Tu, C., **Lederer, A. R.**, [*and 16 others*], Steinmetz, L. M. **Nature Biotechnology** (2018): *10.1038/nbt.4137*

Contributions: analyzed CRISPR-perturbation sequencing data; revised manuscript.

Teaching Experience

Single Cell Transcriptomics: Get Started, BC2 Conference, Basel CH 09/2023
Chosen to organize a one-day tutorial at an international computational biology conference
Single Cell Genomics for Beginners, University of Minho, Portugal (1-week intensive course)

04/2022

Invited to teach an adapted version of an EMBO course to local Portuguese university students

EMBO Single Cell Genomics: Get Started, Braga Portugal (1-week intensive course) 08/2021

Developed course material and taught PhD students the fundamentals of scRNA-seq analysis

Single Cell Biology, EPFL (Profs. Giovanni D'Angelo and Bart Deplancke, 140 hours) 2020 - 2022

Designed course project, hosted exercise sessions, wrote and graded exams

Dynamical Systems in Biology, EPFL (Prof. Felix Naef, 60 hours) 2020

Genetics and Genomics, EPFL (Prof. Bart Deplancke, 30 hours) 2019

Honors and Awards

Poster Prize, Basel Computational Biology Conference 09/2023

Conference Travel Grant, Centre for Genome Regulation 03/2022

NSF Graduate Research Fellowship, National Science Foundation *(declined)* 04/2018

DAAD Graduate Research Fellowship, German Academic Exchange Service 10/2017

Brackenridge Research Fellowship, University of Pittsburgh 08/2016

The Allied Genetics Conference Poster Award, Genetics Society of America 07/2016

Undergraduate Travel Grant, Genetics Society of America 07/2016

Samuel D. Colella Summer Fellowship, University of Pittsburgh 05/2016

Office of Undergraduate Research Award, University of Pittsburgh 08/2015

Chancellor's Undergraduate Research Fellowship, University of Pittsburgh 01/2015

HHMI Summer Fellowship, University of Pittsburgh 05/2014

Four-Year Full Tuition Merit Scholarship to the University of Pittsburgh 08/2013

Selected Conferences

Chan Zuckerberg Institute Annual Meeting (San Diego, USA – **Talk**) 11/2023

Single Cell Genomics 2023 (Engelberg, Switzerland – **Poster**) 10/2023

2nd EDCB PhD Symposium (Lausanne , Switzerland – **Event Organizer** and **Poster**) 10/2023

Basel Computational Biology Conference (Basel, Switzerland – **Talk**/**Poster**) 09/2023

Research in Computational Molecular Biology (RECOMB) (Istanbul, Turkey – **Talk**/**Poster**) 04/2023

Short talk at RECOMB-seq; poster presented at both satellite and main events

Single Cell Genomics 2022 (Utrecht, Netherlands – **Poster**) 10/2022

1st EDCB PhD Symposium (Lausanne, Switzerland – **Event Organizer** and **Poster**) 09/2022

Single Cell Genomics Symposium, Centre for Genomic Regulation (Barcelona, Spain – **Poster**) 03/2022

The Allied Genetics Conference, Genetics Society of America (Orlando, USA – **Talk**/**Poster**) 07/2016

Extracurricular Activities

Student Representative for the EDCB Doctoral Program 07/2020 – 07/2023

Fostered a doctoral student community; organized the first and second editions of a PhD symposium comprising of invited speakers and from across academia, industry, and science communication.

Cellist, University of Pittsburgh Symphony Orchestra 09/2013 - 04/2017