A single cell transcriptomic view of embryonic development and evolution

A dissertation presented
by
James Alexander Briggs
to
The Division of Medical Sciences

In partial fulfillment of the requirements
for the degree of
Doctor of Philosophy
in the subject of
Biological and Biomedical Sciences

Harvard University
Cambridge, Massachusetts
May 2019
A single cell transcriptomic view of embryonic development and evolution

Abstract

A major goal of developmental biology is to define how the cell types of adult organisms emerge from the naive cells of the early embryo. This has been challenging because no methods have been available to comprehensively map the transient intermediate states of differentiation and the detailed molecular processes that occur during cell fate choices. We develop novel approaches for this purpose based on droplet-microfluidic single cell RNA-sequencing (“scRNA-seq”). We show that by recording the transcriptomes of hundreds of thousands of cells in parallel, over multiple points in time, and using methods in high-dimensional data analysis, it is possible to track differentiation from a pluripotent state into many distinct end points simultaneously. The resulting maps provide exquisite detail on the molecular dynamics of differentiation.

Chapters 1-3 focus on mapping fate choices in embryos, and between species. We begin by optimizing experimental methods for scRNA-seq profiling of frog embryos. This allows us to collect a timeseries of >130,000 cell transcriptomes spanning the first 24 hours post-fertilization. We develop computational approaches to comprehensively annotate this dataset by defining all the cell types observed and reconstructing developmental sequences. Next, we compare the resulting differentiation map from frogs with a similar map from fish embryos. We ask how cell types change in gene expression across evolution, and how their developmental histories vary. Finally, we begin to ask systematic questions about patterns of cell fate choice
including how cells transition from continua into discrete expression states, how overlapping gene expression refines at fate choices, and how transcription factors are reused across tissues.

Chapters 4-5 apply similar approaches to identify principles on how artificial systems differentiate in culture. We test whether cell types are sensitive to the methods used to generate them. We show that differentiation is path-independent by identifying two trajectories leading to the same terminal state. And we infer and experimentally test a dose-dependent fate choice mechanism in human organoids.

Overall our results establish general tools that allow data-driven dissection developmental programs, how they change across species, and how they can be engineered to produce cell types of interest in a dish.
Contents

0  Introduction ......................................................... 1
  0.1  The diversity of cells in metazoans  .................................... 2
  0.2  Classifying cell types: concepts and methods  .................................. 3
  0.3  Differentiation of cells in the embryo ........................................ 5
  0.4  Major events in the early embryo ........................................... 6
  0.5  Key molecular events in early cell fate choice  .................................. 9
  0.6  Evolutionary adaptations to embryo development  .......................... 15
  0.7  Differentiation in vitro .................................................. 18
  0.8  Summary: setting the stage for a single cell transcriptomic view of development 20

1  Mapping out embryonic cell states .................................. 24
  1.1  Chapter overview ...................................................... 25
  1.2  Development of a dissociation media for Xenopus embryos ............... 27
  1.3  Single-cell RNA-seq of whole developing embryos  .......................... 28
  1.4  Reconstructing developmental cell state transitions  .......................... 30
  1.5  Assessing pluripotency retention during neural crest development  ........ 34
  1.6  Discussion ............................................................ 38
  1.7  Materials and methods .................................................. 39

2  Comparisons across species ......................................... 51
  2.1  Chapter overview ...................................................... 52
  2.2  Alignment of frog and fish cell state trees ................................... 53
  2.3  Changes in lineage topology across species .................................... 55
  2.4  Properties of conserved genes ............................................ 59
  2.5  Discussion ............................................................ 61
  2.6  Materials and methods .................................................. 62

3  Conserved patterns of cell fate choice ............................ 64
  3.1  Chapter overview ...................................................... 65
  3.2  Appearance of discrete gene expression states .................................. 66
  3.3  Multilineage gene expression priming during fate choices ................. 66
3.4 Re-use of developmental transcription factors ........................................ 70
3.5 Discussion ......................................................................................... 73
3.6 Materials and methods ...................................................................... 73

4 Differentiation plasticity of stem cells .................................................. 77
4.1 Chapter overview ............................................................................. 78
4.2 Dissection of two MN differentiation protocols by scRNA-seq ............. 79
4.3 Dynamics of MN generation versus off-target differentiation .......... 83
4.4 The DP differentiation trajectory lacks Nkx6-1+/Olig2+ intermediate states 84
4.5 Bifurcation and convergence of the DP and SP trajectories ................. 86
4.6 DP transitions through an novel intermediate state .......................... 88
4.7 DP and SP end states closely match MNs from embryos .................. 90
4.8 DP MNs have physiological morphology and function .................... 91
4.9 Discussion ....................................................................................... 95
4.10 Materials and methods .................................................................. 97

5 Reverse engineering fate choice mechanisms .................................... 111
5.1 Chapter overview ........................................................................... 112
5.2 Gata6 overexpression drives multi-tissue organoid differentiation ...... 113
5.3 Mapping cell differentiation in GATA6-organoids by scRNA-seq ....... 115
5.4 GATA6 expression level correlates with early fate choice ................ 117
5.5 Manipulating Gata6-dose to tune organoid cell-type mixture ............ 118
5.6 Discussion ....................................................................................... 119
5.7 Materials and methods .................................................................. 121

6 Conclusions and outlook .................................................................... 127
6.1 Chapter overview ........................................................................... 128
6.2 Contributions of this thesis ............................................................... 128
6.3 Outlook ......................................................................................... 132

Appendix A Supplemental material ....................................................... 139
A.1 Chapter 1 supplemental material ...................................................... 140
A.2 Chapter 2 supplemental material ...................................................... 153
A.3 Chapter 4 supplemental material ...................................................... 156

Appendix B A review of lncRNA mechanisms in the nervous system .... 165
B.1 Abstract ......................................................................................... 166
B.2 Introduction ................................................................................... 166
B.3 LncRNAs are abundant and precisely regulated in mammalian nervous systems 167
B.4 Evolutionary conservation of lncRNA loci ...................................... 168
B.5 Molecular mechanisms of lncRNAs in nervous system development .... 169
B.6 Molecular mechanisms of lincRNAs in neuronal plasticity ........................ 175
B.7 Implicating lincRNAs in neural disease pathogenesis ............................... 180
B.8 LincRNAs in human brain evolution ...................................................... 185
B.9 Conclusion ...................................................................................... 191

REFERENCES ................................................................. 222
<table>
<thead>
<tr>
<th>Number</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.1</td>
<td>Variation in metazoan cell types.</td>
<td>3</td>
</tr>
<tr>
<td>0.2</td>
<td>Cell types in tissues: Tricoplax and the mouse retina.</td>
<td>4</td>
</tr>
<tr>
<td>0.3</td>
<td>The anatomy of early Xenopus development.</td>
<td>7</td>
</tr>
<tr>
<td>0.4</td>
<td>Molecular events underlying symmetry breaking in the Xenopus egg.</td>
<td>11</td>
</tr>
<tr>
<td>0.5</td>
<td>Long range signaling gradients spatially organize cell types in the</td>
<td>13</td>
</tr>
<tr>
<td></td>
<td>Xenopus embryo.</td>
<td></td>
</tr>
<tr>
<td>0.6</td>
<td>Spemann’s organizer: axis twinning and molecular signals.</td>
<td>14</td>
</tr>
<tr>
<td>0.7</td>
<td>Transcription factor cross-antagonism can drive segregation of discrete expression states.</td>
<td>14</td>
</tr>
<tr>
<td>0.8</td>
<td>The hourglass model of evolutionary adaptations in embryogenesis.</td>
<td>17</td>
</tr>
<tr>
<td>0.9</td>
<td>Stem cells for regenerative medicine and understanding differentiation plasticity.</td>
<td>19</td>
</tr>
<tr>
<td>0.10</td>
<td>Research strategy: reconstructing developmental manifolds by scRNA-seq.</td>
<td>21</td>
</tr>
<tr>
<td>1.1</td>
<td>Dissection of early Xenopus tropicalis development by scRNA-seq.</td>
<td>29</td>
</tr>
<tr>
<td>1.2</td>
<td>Inference of developmental cell state transitions from gene expression similarity.</td>
<td>32</td>
</tr>
<tr>
<td>1.3</td>
<td>scRNA-seq detects early transcriptional events during specification of embryonic cell states.</td>
<td>34</td>
</tr>
<tr>
<td>1.4</td>
<td>The Xenopus cell state tree is a powerful gene expression resource.</td>
<td>35</td>
</tr>
<tr>
<td>1.5</td>
<td>Assessing the retention of pluripotency during neural crest development.</td>
<td>37</td>
</tr>
<tr>
<td>2.1</td>
<td>Frog and fish single cell maps.</td>
<td>53</td>
</tr>
<tr>
<td>2.2</td>
<td>Alignment of frog and fish cell state trees.</td>
<td>54</td>
</tr>
<tr>
<td>2.3</td>
<td>Topological rearrangement of hatching gland differentiation trajectory across species.</td>
<td>56</td>
</tr>
<tr>
<td>2.4</td>
<td>Expression similarity does not predict cell type orthologs.</td>
<td>58</td>
</tr>
<tr>
<td>2.5</td>
<td>Properties of conserved genes.</td>
<td>60</td>
</tr>
<tr>
<td>3.1</td>
<td>Dynamics of discrete cell state appearance in Xenopus in scRNA-seq timecourse.</td>
<td>67</td>
</tr>
<tr>
<td>3.2</td>
<td>Refinement of promiscuous multilineage gene expression during early embryonic fate choices.</td>
<td>69</td>
</tr>
<tr>
<td>3.3</td>
<td>MLP genes are enriched for TF-associated GO terms.</td>
<td>70</td>
</tr>
<tr>
<td>3.4</td>
<td>TF reuse is pervasive in vertebrate development.</td>
<td>72</td>
</tr>
<tr>
<td>4.1</td>
<td>Summary of research questions.</td>
<td>79</td>
</tr>
<tr>
<td>4.2</td>
<td>Dissection of DP and SP motor neuron differentiation strategies using InDrops single cell RNA sequencing.</td>
<td>82</td>
</tr>
</tbody>
</table>
DP and SP induce distinct differentiation paths to the motor neuron state. 87
During DP cells transition through an abnormal intermediate state with a forebrain gene expression defect. 89
Both DP and SP differentiation trajectories approach the transcriptional state of primary MNs (pMN), but DP does so with higher precision. 92
Validation that DP cells become functional MNs despite their abnormal differentiation trajectory. 94

Gata6-driven differentiation of multi-tissue human organoids. 114
Mapping Gata6-organoid differentiation by scRNA-seq. 116
Gata6 expression level correlates with output cell fate. 118
Doxycycline dose allows modulation of Gata6 levels. 119
Manipulating Gata6-dose to tune organoid cell-type mixture. 120

Acorn worm scRNA-seq timecourse. 136
Whole embryo scRNA-seq timeseries data across the tree of life. 137

Optimization of dissociation and inDrops procedures for Xenopus. 141
scRNA-seq sensitivity analysis. 142
Example of nested clustering. 143
Computational cross-validation of stage 22 tSNE-DBSCAN clustering by comparison to SPRING-spectral clustering. 144
Structure of tSNE-DBSCAN clusters is preserved on an independent visualization of the underlying scRNA-seq data. 145
(continued) 146
Voting outcomes during ancestor assignment across timepoints. 147
(continued) 148
Marker gene expression on the cell state tree. 149
Comparison of cell state tree structure to alternative cell state coarse graining algorithm. 150
Differentiation asynchrony in the somites. 151
Expression at stage 11 of individual genes proposed to constitute a retained blastula pluripotency circuit in neural crest cells. 152
Proportion of cell states and total cells matched between Xenopus and Zebrafish cell state trees. 153
Changes in germ layer proportions between Xenopus and Zebrafish. 154
Global gene expression similarity of matched Xenopus and Zebrafish cell states. 155
tSNE visualization of marker genes for DP subpopulations. 157
(continued) 158
tSNE visualization of marker genes for SP subpopulations. 159
(continued) 160
| A.16 | qPCR validation that MN gene expression in DP follows the expected dynamics and does not involve Olig2 induction. | 161 |
| A.17 | Joint visualization of DP and SP differentiation paths by tSNE. | 162 |
| A.18 | Cell-cycle-associated gene expression decreases as cells enter the DP and SP intermediate states and then progressively mature. | 163 |
| A.19 | Heterogeneity within the Mnx1+ E13.5 primary motor neuron population. | 164 |
| B.1  | Regulatory mechanisms of IncRNAs. | 170 |
| B.2  | Summary of IncRNA roles in neuronal development and their associated complexes. | 176 |
| B.3  | Molecular mechanisms of IncRNAs in synaptic plasticity. | 181 |
This thesis is dedicated to my family in Australia.
Acknowledgments

I would like to thank all the talented people around me in the Department of Systems Biology at HMS, who by their daily brilliance inspire and teach me to pursue research to an ever higher standard. I feel incredibly lucky to have had the opportunity to learn from such wonderful people with such diverse backgrounds, especially those different to my own in computer science, physics, and applied mathematics. The process of sharing an interest in a biological problem, and working together, but coming at it from the perspective of entirely different backgrounds has been I think both tremendously powerful and wonderfully educational. In this department and pair of labs I have found everything I had hoped I would find in moving across the planet for graduate school.

Special thanks go to my lab mates Sam Wolock and Caleb Weinreb, who by never failing to share their time and knowledge have played a special role in educating me to become a better interdisciplinary scientist. Thanks also to Dan Wagner, Victor Li, Patrick Fortuna, and Jessica Gray who collaborated with me to work across multiple experimental systems. And to Naren Tallapragada who as my bench mate and one time roommate has provided close and valued friendship over the past five years.

Most of all I would like to thank my advisors Marc Kirschner and Allon Klein for their close mentorship, generosity, and direction. I have been extraordinarily privileged to work under these two excellent scientists, but also wonderful people. Marc comes from a background in biology (it is hard to define his contributions more narrowly) and has always encouraged me to think big and to pursue problems that I think are important. There is seemingly no topic on which Marc doesn’t possess a deep knowledge and the ability to find interesting questions. And working at the bench together I have learned tremendously from his quick intuition for the key issues in experimental science. Allon is from a very different background, with training in theoretical physicist, and provided wonderful complementary mentorship during my transition toward more quantitative biology in graduate school. Aside from his obvious brilliance, and contributions to our science, Allon has an unusual capacity to distill problems into a very clear conceptual form, and to communicate this concretely and intuitively; these traits have made him a wonderful teacher to me during graduate school.

Finally, I would like to thank Katherine, for constantly supporting me over the past four years as my single closest companion and friend. Whether taking weekend trips to escape the city and clear our minds, or simply hanging out in the evenings after a day in lab, you have provided a constant source of fun, balance, and peace that I can’t imagine being without. And to my friends outside of lab (you know who you are) life wouldn’t be nearly as good without you; for knowing you, I am extremely grateful.
0

Introduction
0.1 The diversity of cells in metazoans

Cells are the fundamental building blocks of multicellular organisms. Across metazoa they appear in a remarkable diversity of cell types, each with unique molecular compositions and functions. The partitioning of functions among specialized cell types in tissues is a strategy that has been used by evolution to build organs that can carry out complex physiological tasks.

Cells were first observed in cork by Robert Hooke in the 1650’s, following the invention of the first microscopes (Fig. 0.1A)\textsuperscript{112}. By the 1670s-80s Antoni van Leeuwenhoek observed the first protozoa and bacteria. It took until 1838 to realize that animals too were made of cells, when Schleiden and Schwann formulated cell theory: the idea that all living things are composed of cells or their products\textsuperscript{243,244}. We now appreciate that cells contain a rich assortment of internal machineries in the form of organelles and differ from one another across the body, partitioning into a variety of identifiable “types”.

Across metazoa, there is a staggering diversity of cell types. Cells range in morphological and compositional complexity from extremely simple to exquisitely complex. Red blood cells are little more than a bag of hemoglobin, specialized to shuttle oxygen around the body. By contrast, Purkinje neurons of the cerebellum form tens of thousands of arborized dendritic spines that synapse with hundreds of thousands of granule cells and coordinate the coherent motor functions of the animal (Fig. 0.1B). Each type of cell is specialized to perform a particular function with maximum effectiveness.

The partitioning of functions across many specialized cell types allows complex physiology in higher metazoans. Simple metazoa such as Trichoplax have just 6 identifiable cell types (Fig. 0.2A), including dorsal and ventral epithelial cells, lipophils, fiber cells, gland cells and crystal cells\textsuperscript{245,246}. The cells individually resemble cells in humans. Yet Trichoplax lead simple lives, moving around slowly by the power of ciliated epithelial cells on its ventral side, and passively absorbing nutrients - primarily in the form of microbes - from the environment. By contrast humans have been estimated to have as many as 2,000 cell types - though the full number may be much higher\textsuperscript{390} - each arranged carefully to form organs that support our comparatively
Fig. 0.1: Variation in metazoan cell types. (A) Drawing of the first cells ever observed in cork by Robert Hooke, 1665. (B) Two extreme examples of cell complexity: a Purkinje neuron in green (i) compared to a red blood cell (ii). Scale bar = 50μm. Zoom in panel shows one of the 100,000 synapses that each Purkinje neuron forms with different granule cells to control motor functions. Purkinje neuron images adapted from Boris Barbour lab. RBC image from cgtrader.com.

more complicated lifestyles.

The retina alone contains 39 distinct cell types (Fig. 0.2B), each with a unique molecular composition and specialized function. Without these many specializations it would be impossible to build a functional retina. Varieties of rod and cone cells perceive light at specific wavelengths. Bipolar, ganglion and amacrine cells form specific connectivities, respond to the right inputs, and relay the correct outputs. And Muller glial provide essential metabolic support. Each cell type plays a role in converting the light that enters our eye neural representation of the outside world, perceived by our brains in color and three dimensions.

0.2 Classifying cell types: concepts and methods

Cells are typically classified into types by partitioning populations of cells into groups that are defined by a unique morphology, function, molecular composition, or ontogeny. Taxonomies of cell types made by each criteria are not guaranteed to be consistent, and debate can surround the most appropriate definition of cell type. Nonetheless they offer useful operational frameworks to discuss cell diversity. The more detailed the information that can be obtained on each cell, the richer and more precise cell type classifications can
A variety of techniques have historically been used to define differences between cells and classify them into types. Early classifications relied on differences in: morphology; staining with dyes (e.g. hematoxylin); cell density under centrifugation; cell behavior (e.g. neuronal firing patterns or circus movements in culture); or developmental potential (e.g. CFU versus BFU tests of blood progenitor clones). Later, molecular methods including antibodies, in situ staining, and genetic labeling were developed to distinguish cells based on differences in specific ‘marker genes’. Molecular approaches have proved especially useful due to their convenience, and ability to distinguish even morphologically similar cells with different functions or ontogeny. Conversely, cells with a specific function can generally be defined by a unique molecular profile.

The methods used to record molecular differences between cells have become increasingly sophisticated over time. Early chromogenic detection methods could detect one or at most two distinct labels in tissues. Fluorescent tagging techniques allowed multiplexing of up to five labels at cellular or even subcellular resolution. And cell sorting methods such as FACS allow rapid profiling of fluorescent markers on many individual cells. Mass cytometry methods have recently extended the multiplexing capacity of cell sorting to allow detection of tens of protein markers per cell. And very recently, single cell RNA sequencing (‘scRNA-seq’) has allowed measurement of the entire transcriptome in single cells at scale. While mass-cytometry
remains challenging to implement, scRNA-seq has been deployed extremely broadly over the past three years due to its accessibility, utility, and affordability.

The advent of scRNA-seq is revolutionizing our ability to define cell diversity in complex tissues. In contrast to previous approaches that relied on pre-selected marker genes, or population averaged measurements (e.g. by bulk RNA-seq), scRNA-seq uniquely allows the systematic and unbiased characterization of gene expression states in complex populations. By using a data-driven and high-dimensional approach to enumerate cell states, scRNA-seq avoid the pitfalls of marker genes such as overlooking population substructure, and of bulk methods which can introduce averaging artifacts. In practice scRNA-seq can resolve subtypes of cells within a group previously thought to be a single population on the basis of marker genes. And by computationally parsing cells into pure populations, scRNA-seq can provide transcriptional profiles of cell states that are free of the population averaging artifacts that normally arise from bulk RNA sequencing of heterogeneous cell populations. Since scRNA-seq requires no prior knowledge about the composition of a tissue to discover its constituent cell states, it is uniquely able to support rapid characterization of poorly studied tissues.

Currently, scRNA-seq is the only single cell “omics” technology mature enough to apply broadly to characterize cell diversity in tissues. However, not all differences between cell types (e.g. post-translational modifications of proteins) are guaranteed to be visible at the level of transcription. Eventually, additional measurement modalities such as proteomics, metabolomics, and chromatin state profiling approaches are therefore expected to emerge and further enrich our understanding of cell diversity.

0.3 Differentiation of cells in the embryo

The cell types of the adult form from naïve pluripotent cells found in the early embryo by a process of differentiation. We now well appreciate that differentiation is a process that occurs in stages, and is largely driven by transcriptional changes. This was not the case a century ago. In the early 19th century the “preformation-
sts” argued that the egg was a very complex cell that concealed the complexity of the adult. This view was only overturned by the “epigeneticist” Karl Ernst von Baer in 1828, who showed that the complexity of the embryo emerges in stages. It took until the early 20th century until geneticists including Thomas Hunt Morgan showed in Drosophila that developmental processes were under the control of genes. With the development of molecular biology through the 50s and 60s it became possible to study how genetic changes propagate through functional RNA and protein intermediates to control the behavior of cells as they differentiate.

The big question became whether we can record the sequence of molecular changes that occur in cells during differentiation, and then can we understand how these changes control cell fate choices. Enormous progress has been made in this area. Although there is too much work to comprehensively survey, the next two sections review some of the key macroscopic events that occur in early embryos, and the underlying molecular events. The approaches taken can be summarized as usually addressing one of two powerful questions: (1) can we identify developmental relationships between cells, in order to link states over time? (2) can we obtain comprehensive measurements on the state of purified cell populations at each time point? These approaches are powerful but are work intensive.

0.4 Major events in the early embryo

While changes in gene expression occur within each cell individually in vertebrates, differentiation occurs as part of developmental patterning, which is a coherent process across cells in space and time. We look at several organisms in the chapters that follow but use the frog embryo here as an example. The major events that take place during the first day of development are: fertilization, cleavage, the mid-blastula transition (MBT), germ layer specification, gastrulation, neurulation, and primary axis elongation (Fig. 0.3). The end result is an embryo organized into rudimentary body plan, with ~250,000 cells partitioned into tens of distinct cell types, each specified in the correct numbers and at the correct spatial locations.
Fig. 0.3: The anatomy of early Xenopus development. Cartoons showing the spatial organization of the embryo at the major stages of early Xenopus development. Cell types are colored by germ layer and tracked through the morphogenetic movements of gastrulation. Figure originally from De Robertis et al. 61.
The frog oocyte is a radially symmetrical cell and is divided into a pigmented animal (top) and vegetal (bottom) domain. Fertilization by sperm triggers a rotation of the animal hemisphere, towards the sperm entry site, which can be recognized by the formation of a grey crescent on the future dorsal side of the embryo. The embryo then undergoes a series of 12 rapid cleavage cycles (divisions without growth), which divide the egg into a blastula of 4,096 cells by 5 hours post fertilization (hpf). The blastula stage was originally defined by the formation of an internal cavity known as the blastocoe (Fig. 0.3). At the 4,096 cell stage the embryo undergoes what is known as the midblastula transition, which involves the activation of zygotic genome transcription, cell cycle lengthening, and the beginning of cell motility. Shortly after MBT, by 9hpf, the three germ layers become defined. The ectoderm forms the roof of the blastocoe. The endoderm forms from the vegetal blastomeres. And the mesoderm forms in a ring between the ectoderm and endoderm.

Lewis Wolpert famously said: “It is not birth, marriage, or death, but gastrulation which is truly the most important time in your life”. Indeed the complex choreography of morphogenetic movements that occur during gastrulation pattern the embryo into a rudimentary body plan. Gastrulation begins at 10hpf with the involution of mesoderm through the dorsal blastopore l. The involuting mesoderm forms a layer of tissue between the endoderm and ectoderm. As gastrulation proceeds, the blastopore forms a ring around the entire embryo, and migrates downward, pulling the ectoderm with it until cells of the ectoderm eventually cover the whole surface of embryo. Ultimately the embryo is divided into three layers and adopts its true anterior-posterior and dorsal-ventral axes (the coordinate system rotates 90° with the animal pole becoming anterior of the embryo). Gastrulation is also a period of extensive patterning by intercellular communication. The ectoderm, mesoderm, and endoderm are all subdivided into multiple distinct progenitor states by inductive interactions as they migrate through the embryo. We will discuss the molecular details of these events in the following section.

Upon completion of gastrulation, around 14hpf, the embryo forms a neural plate on its dorsal side, defining the neurula stage. Over the next few hours the neural plate folds along the anterior posterior axis, fuses,
and becomes internalized to form the neural tube. The embryo also elongates along its anterior posterior axis due to the intercalation of internal mesodermal cells, and continues to pattern. The neural plate is divided into forebrain, midbrain, hindbrain, and spinal cord compartments. The eye fields are specified in anterior ectoderm. The skin is divided into ciliated cells, ionocytes, goblet cells, and small secretory cells. The notochord forms and collagenizes. The somites are budding to form the future vertebral columns and axial muscle. And the endoderm is divided into foregut and hindgut domains of the future digestive system. By 24hpf, the progenitors of most major organs are established and the embryo is described as tailbud stage, due to the formation of the tailbud, which will subsequently proliferate to form the extended tail of the tadpole. In total cells in the embryo transition from a single pluripotent state in the early blastula, to three germ layers, then tens of early progenitor domains in a late gastrula, and over fifty progenitor compartments by tailbud stages. As cells diversify, they gradually restrict their fate potential until they acquire the properties of specific mature cells.

In the thesis we will also be considering zebrafish (Danio rerio) and Acornworm (Saccoglossus kowalevskii). Zebrafish is separated by 435 My evolution from frog, and undergoes many of the same stages of development. In Acornworm, by contrast, significant differences in development have accumulated over 684 My evolution. Acorn worms have a dorsal-ventral axis inversion, with a ventral nerve cord and dorsal gut. They also lack a centralized nervous system, instead displaying a nerve net across the surface of the animal. Nonetheless, many developmental events (even the patterning of the structurally distinct nervous system) appear to be conserved making for an interesting comparison.

0.5 Key molecular events in early cell fate choice

There is now some understanding of the major molecular processes involved in forming the different germ layers, followed by appearance of increasing numbers of cell types. Here we introduce the key mechanisms underlying the symmetry breaking of the egg, the patterning of the blastula into discrete germ layers, and the
formation and function of Spemann’s organizer. We briefly summarize how patterning continues through gastrulation to generate the full diversity of cell types. We will focus again on frog as an example, but refer to other organisms where useful to illustrate the general principles.

0.5.1 Symmetry breaking of the egg: the origin of spatial pattern

The first patterning event of development is the breaking of dorsal-ventral symmetry in the egg. In frog, this occurs upon fertilization of the oocyte. Fertilization triggers polymerization of microtubules in the direction opposite of the sperm entry site. These microtubules carry microvesicles loaded with the dorsal determinants Frz7 and Dsh (that are maternally deposited in the vegetal pole of the egg) towards the future dorsal side of the embryo (Fig. 0.4)\textsuperscript{216,293,271}. There, Frz7 and Dsh then inhibit GSK\textsubscript{3}, which stabilizes B-catenin. B-catenin then accumulates and translocates to the nucleus on the dorsal side of the embryo, where it functions as a transcription factor. In the nucleus it heterodimerizes with Tcf-3, and activates dorsal determinant genes including the homeobox genes siamosis and xtwin, as well as the secreted factors nodal-related-3, and cerberus, thus specifying dorsal cell identity\textsuperscript{195,30,54}.

0.5.2 Patterning the blastula: organization by long-range signaling gradients

The patterning of the blastula into ectoderm, mesoderm, and endoderm, and their subtypes along the dorsal-ventral axis, illustrates how long range signaling gradients pattern tissues in embryos. The frog blastula is patterned in two main stages. First, a complex interplay between maternally deposited vegetal factors - which establish animal-vegetal asymmetry - and the newly established dorsal-ventral B-catenin axis (Fig. 0.5; left) establish the Spemann’s organizer on the dorsal side of the embryo. Second, Spemann’s organizer secretes a collection of signaling molecules that form gradients across the embryo, leading to the activation of key lineage-specific transcription factors at defined spatial locations.

The maternal animal-vegetal axis: VegT (a transcription factor) and Vg1 (a secreted signaling molecule)
Fig. 0.4: Molecular events underlying symmetry breaking in the Xenopus egg. (A) Fertilization triggers microtubule polymerization, and transport of microvesicles containing the dorsal determinants Frz7, Dsh, and Wnt11 to the future dorsal side of the embryo. These factors induce dorsal cell type specification by inhibiting GSK3 and thus stabilizing B-catenin, which then translocates to the nucleus to activate dorsal determinant genes. LiCl can be used to experimentally induce a dorsal axis by directly inhibiting GSK3. (B) Blocking microtubule polymerization by UV, nocodazole, or low temperature causes failure of dorsalization, showing the requirement of microtubule polymerization for symmetry breaking. Figure originally from De Robertis et al. 
are maternally deposited at the vegetal pole as mRNAs (for a review of other factors see). They are translated during cleavage. At MBT, they activate endodermal specifier genes including Sox17, Mixer, Gata-4, 5 and 6 in vegetal cells, as well as several nodal related signaling molecules (the 5 Xnr’s), and Wnt8,125,37. These signaling molecules are then secreted along with Vg1 to induce mesodermal specifier genes including T, and Xbra in the overlying cells.28,146,3. Animal pole cells that do not receive mesoderm inducing signals adopt an ectodermal fate defined by expression of Sox2.103.

The dorsal-vental axis: While the vegetal signals are initially specifying endoderm or mesoderm fates after MBT, the dorsal-ventral B-catenin axis is acting simultaneously to pattern the endoderm, mesoderm, and ectoderm into subtypes depending on their spatial location. Endoderm cells near the dorsal B-catenin signal (designated as the Nieuwkoop centre) express higher levels of nodal related signaling molecules, including Vg1, as compared to the ventral endoderm.1 As a result dorsal endoderm induces dorsal mesoderm fates (Spemann’s organizer, and future tailbud), marked by genes including cerberus (the Bmp and Wnt antagonist), while ventral endoderm induces ventral mesoderm fates (future blood and lateral plate).3. The dorsal B-catenin signal also influences mesodermal patterning by altering nodal target gene activation at the level of genome binding by Smad2, e.g. at the Gsc promoter, in a separate cell intrinsic mechanism.91. In the ectoderm, Bmps induce epidermal fate.106 Nuclear B-catenin suppresses responses to Bmp activity to promote neural fate in dorsal cells.25,12,16.

Establishing Spemann’s organizer: a prototype signaling center The dorsal mesoderm of the late blastula forms a tissue known as Spemann’s organizer, which is perhaps the most famous signaling center in embryology. Hans Spemann showed in 1924 that transplantation of the dorsal mesoderm to the ventral side of a donor embryo could induce a second body axis, resulting in an embryo with two separate heads fused at the trunk.268 (Fig. 0.6A). Since that initial experiment, discovering the molecular basis of the organizers inductive function was a kind of holy grail of embryology. In the 80s and 90s, the tools of molecular biology finally made this possible. It was shown that the organizer secretes a cocktail of inhibitory molecules including Noggin, Chordin, Cerberus, Frzb-1, Crescent and Dickkopf (DKK), among others (Fig. 0.6B). More recently it
was discovered that an opposing signaling center also exists on the ventral side of the embryo that secretes factors including CV-2, Bmp4, Tsg, Xlr, Bambi and Sizzled (Fig. 0.6B). Together with the pre-existing animal-vegetal and dorsal-ventral signaling gradients, these molecules collaborate to activate transcription factors that specify the three germ layers, and their subtypes, in the correct spatial locations. It is now appreciated that factors in the dorsal and ventral signaling centers interact to allow pattern scaling in the frog embryo - i.e. such that an embryo experimentally reduced to half size will form all of the normal tissue domains but on a smaller overall length scale$^{10,18}$.

0.5.3 PATTERNING THE BLASTULA: SEGREGATION OF DISCRETE EXPRESSION STATES

Cell state specific transcription factor expression defines the germ layers of the blastula. Although the activation of these factors is spatially organized by signaling gradients, the precision of domain formation is not perfect. The initial activation of transcription factors can overlap in single cells$^{106}$, which then require additional mechanisms to refine their fate into a single state and to form sharp and coherent spatial boundaries. Three major mechanisms contribute: cross-antagonism of specifier genes$^{28,92}$ (Fig. 0.7), community effects$^{95}$, and cell sorting$^{39,51}$. However, this phenomenon remains poorly understood in the early embryo.
Fig. 0.6: Spemann’s organizer: axis twinning and molecular signals. (A) The Spemann-Mangold experiment. A control embryo (top) is shown next to: (bottom left) a gastrula stage host embryo, indicating the transplanted organizer tissue (white tissue at bottom of blastopore); and (bottom right) the twin axis tadpole that results from the transplant. (B) Molecules secreted by the ventral center and Spemann organizer. Panels originally from De Robertis et al. 60.

Fig. 0.7: Transcription factor cross-antagonism can drive segregation of discrete expression states. (A) Schematic illustration of two transcription factors (top) that: i) mutually inhibit each other; ii) self activate; iii) activate distinct sets of downstream genes; and iv) inhibit the downstream genes activated by the other factor. These four properties can help to generate mutually exclusive expression of different transcriptional states. (B) Schematic of a differentiation potential landscape. Cells are depicted as balls and colored by their expression levels of the two competing regulators from (A). Progenitor cells at the top of the landscape have multilineage expression and fluctuate between states with probabilities proportional to their expression levels of each factor. Eventually they fluctuate out of the progenitor basin into one of two distinct end states, illustrating a cell state bifurcation. Panels adapted from Graf et al. 92
0.5.4 Patterning in gastrulation: onset of cell migration

In the blastula, cell movement is essentially negligible. Patterning operates in a static spatial coordinate system. This situation changes dramatically during gastrulation. Extensive cell migrations lead to each cell traversing a range of signaling environments over time. These environments are specified by both long range diffusible signals, but also cell-cell interactions as populations move past one another. Cells are therefore patterned by highly dynamic signaling histories. The full interactions occurring during this period are too numerous to discuss comprehensively, but the events patterning the involuting mesoderm illustrate the principle.

0.5.5 Patterning the epidermis: organization by local cell interactions

The patterning mechanisms described so far generally drive fields of cells into a similar fate, separated at tissue boundaries. However, cell types can also pattern into precise mixed populations as in the epidermis. Formation of these patterns can occur by cell-cell interaction and reaction diffusion mechanisms.

0.6 Evolutionary adaptations to embryo development

Since the the 19th century it has been appreciated that evolutionary changes across species arise by changes in the development of the embryo$^{31,299,100}$. More recently, our growing knowledge of the full details of the molecular processes regulating development have lead to new insights into evolutionary mechanisms. We now appreciate that adaptations are in fact facilitated by the modular nature of developmental programs$^{39,42}$. Cells types are an example of modular organization, and provide a compartmentalized substrate which evolution can adapt and redeploy to generate new physiological capabilities. However, we currently understand fairly poorly the mechanisms underlying the evolution of cell types.

Evolution of new cell types. All metazoan cell types are believed to descend from a single ancestral cell - the choanocyte$^{9,10}$. This ancient cell is thought to have resembled chaonoflagellates (the nearest unicellular
relative of metazoa), with large polarized flagella, a collar of microvilli, and well-developed endocytosis machineries for nutrient uptake. Genomic evidence indicates that choanocytes already had many of the genes found in diverse cell types of higher metazoa, including for example genes encoding the molecules that assemble to form the synaptic machinery found on neurons in higher metazoa. New cell types are thought to have descended from this ancestral cell by two main mechanisms: functional segregation of exiting genes into distinct types; and the acquisition of new gene functions in some cells of the organism. The evidence for this comes from the fact that genes that were expressed together in choanocytes are found to be expressed selectively across cell types of higher metazoa today. In addition, genome sequences have revealed complex phylogenies of gene families, where ancestral genes in choanocytes have repeatedly duplicated and evolved new functions over time.

In all but the simplest cases we understand fairly poorly the precise developmental mechanisms leading to the formation of new cell states. A fascinating example of what might begin to understand comes from sponges, which are among the closest metazoan relatives of choanoflagellates. In sponges, there are two cell layers - an outer layer and an inner layer. It was recently found that the inner layer of sponge cells form in response to the activation of the transcription factor GATA, associating this factor with an ancient cell type evolution event\(^\text{106}\). Intriguingly, in higher metazoa, GATA family transcription factors specify the endomesoderm germ layer\(^\text{176}\). It has been suggested that this molecular similarity may indicate a common origin between the two layers of sponges, and the germ layers of higher metazoa. As we move further along the evolutionary tree the events driving the formation of new cell types become even more murky.

**Redeployment of cell types.** Across species, new cell types have evolved but existing cell types have also been redeployed across tissues to allow new functions. Ionocytes are one example of a cell type that was ostensibly redeployed in many tissues. They can be found in various tissues among vertebrates, including the gut, airway, or skin, depending on the species. They were originally discovered in the frog epidermis, and later recognized in human airway\(^\text{68,221}\). In both frog skin, and human airway they are specified the same conserved transcription factor, Foxii, supporting the idea that the activation of this master transcription factor in a new
location of the developing embryo is responsible for the appearance of this conserved cell type in multiple tissues. Further support for the idea that master regulators can drive relocation of cell types comes from experiments on the eye development master regulator, Pax6, which is expressed in the eyes of species ranging from flies to humans, spanning 1,000 My of evolution. In experiments where the expression of Pax6 is induced in ectopic body locations in the fly or in frogs, such as the legs or back, whole eyes are observed to develop in full morphology in the ectopic body location\textsuperscript{102,49}.

\textit{Growth and specialization of cell types.} Organisms can adapt without the invention of additional cell types or rearrangement of cell types in two ways. They can change the numbers of cells in a tissue. For example the skin that covers the wings of bats as grown tremendously as the wings evolved to allow flight, but remains composed of the same types of cells. They can also specialize their existing cell types. For example gene modules encoding a specific function, such as cilia formation, can be activated or lost from cell types. Tracking these changes over orthologous cell states across large evolutionary distances has been challenging.

\textit{Phylotypic stage.} Since the 19th century it has been appreciated that animals with very different adult body plans, and even very different eggs, appear more similar in embryonic stages of development, leading to the

\textbf{Fig. 0.8: The hourglass model of evolutionary adaptations in embryogenesis.} Animals with very different eggs and adult body plans appear more similar in an embryonic stage of development known as the phylotypic stage. The molecular underpinning of this phenomenon, indicated by the schematic on the right, are poorly understood. Figure originally from Hu et al\textsuperscript{113}. 

---

idea of a conserved stage of development that was common to all members of a phylum - the phylotypic stage (Fig. 0.8). This idea has gained additional support from molecular studies, which have found striking conservation of gene expression across orthologous cell compartments in the early embryo\textsuperscript{89}. The reason that intermediate stages of development are the most conserved is incompletely understood.

0.7 Differentiation in vitro

The advent of stem cell and cell reprogramming technologies has added practical incentives to study cell differentiation. By controlling cell fate in culture it may become possible to derive replacement cell types or tissues for regenerative medicine. Beyond this practical utility, cell reprogramming technologies have raised interesting conceptual questions about the plasticity of differentiation and provided an accessible experimental system in which to test them.

\textit{Controlling differentiation outcomes}. Since the isolation of the first embryonic stem cells from mice in 1981\textsuperscript{75,73}, and humans in 1998\textsuperscript{281}, and the advent of reprogramming to pluripotency in 2006\textsuperscript{276,275}, enormous progress has been made showing that these pluripotent cells can be driven into desired cell types in culture by addition of the correct growth factor signals\textsuperscript{50}. These cells in culture are useful tools for disease modelling, drug discovery, and are hoped one day to allow replacement of damaged tissues in regenerative medicine (Fig. 0.9A). A major focus of stem cell biology is the invention of new and more efficient methods to direct the differentiation or reprogramming of cells into desired cell types. More recently, work has begun to explore how stem cells can be patterned into three-dimensional ‘organoids’ composed of multiple cell types\textsuperscript{354}.

\textit{Testing principles of cell fate plasticity}. The process of reprogramming is also scientifically interesting because of what it teaches us about differentiation plasticity. After it was first shown in 1980 that induction of MyoD could convert fibroblasts into muscle cells\textsuperscript{18}, it took until 2006 to demonstrate reprogramming to pluripotency by transcription factors\textsuperscript{275}. Since then an explosion of reprogramming recipes have been reported that interconvert a wide range of cell types\textsuperscript{149}. In contrast to the hierarchical differentiation seen
in embryos, these reprogramming approaches indicate that the exact sequence of intermediates defining a lineage may be more flexible. In particular, it appears that mature cell states can be reached through paths that do not involve activation of the intermediate progenitor genes that are essential in embryos. Fibroblasts can be converted into mature neuron phenotypes, seemingly without completely dedifferentiating and retracing the embryonic lineage, as indicated by lack of expression of specific core pluripotency (Oct4, Sox2 and Nanog) and neural progenitor genes (Nestin). Mouse embryonic stem cells (mESC) can also be converted into motor neurons (MNs) by a process that involves overexpression of three transcription factors, Ngn2+Isl1+Lhx3, and that never expresses the neural progenitor transcription factors Sox1 and Olig2. Intensive effort is now being spent to understand whether cell types produced by these artificial pathways are equivalent to the cell types generated by the embryonic pathways, and to understanding the differentiation trajectories followed by cells in reprogramming. Resolving these issues has deep implications for our understanding of cell state plasticity.
0.8 Summary: setting the stage for a single cell transcriptomic view of development

0.8.1 Motivation

Animals are built from cells that appear in nature as specialized types – e.g. muscle, skin, neurons, and bone. This diversity of cell types arises in the embryo by the process of cell differentiation: the fertilized egg divides repeatedly while daughter cells gradually acquire distinct functional, morphological, and molecular characteristics. Since at least the 1800s developmental biologists have asked how cell differentiation occurs with two kinds of motivation: i) to understand how nature builds complex and diverse animal forms, and ii) to learn how we might control these processes to generate cells artificially.

However, the process of increasing cell-type heterogeneity has been difficult to follow comprehensively at the molecular level. A conventional biochemical dissection of the underlying process is complicated by the limited material in the embryo and the heterogeneity of its composition, both of which must be considered together. Many insights have emerged from genetic studies and experimental embryology, but those analyses generally focus on a few transcription factors and/or secreted signaling molecules. Instead, development involves parallel alterations in many different intracellular and intercellular circuits.

There is therefore a need of new methods to measure the complex molecular changes occurring in cells as they differentiate. A complete view of these molecular processes could reveal: new cell subtypes defined by previously unknown marker genes; novel factors associated with cell fate choices; or entirely new processes visible only in the collective fluctuations of many genes.

0.8.2 Research strategy

In this work, we aimed to comprehensively map the transcriptional changes occurring during differentiation across whole embryos using single cell RNA-sequencing. By profiling large numbers of cells from embryos
Fig. 0.10: Research strategy: reconstructing developmental manifolds by scRNA-seq. We aimed to collect single cells from embryos across multiple points in time, measure their transcriptional state by scRNA-seq, and then connect similar states over time into developmental manifolds using high-dimensional data analysis techniques. The resulting ‘map’ would reveal the transcriptional changes leading to the formation of each and every cell type in the embryo.

over time, we reasoned that it may be possible to infer differentiation trajectories into every cell type in parallel by using methods in high-dimensional data analysis (Fig. 0.10). This idea was exciting because the resulting transcriptional map would in principle allow us to read out the genome-wide gene expression dynamics that restrict the identity of each and every cell type in the embryo in silico. It would allow interrogation of branching points in development. And it would allow quantitative comparisons of differentiation across species and with cells in culture. Still, studying differentiation across entire embryos by scRNA-seq presented highly non-trivial challenges both experimentally and computationally.

Two critical developments set the stage for us to attempt this objective. First, Allon Klein - as a post-doc in the Kirschner lab - developed a droplet-microfluidic single cell RNA-sequencing technology that opened the possibility of profiling tens to hundreds of thousands of single cell transcriptomes affordably (compared to hundreds for alternative best in class methods at the time)\textsuperscript{141}. Second, two very recent studies had demonstrated how this scRNA-seq data can allow marker gene free decomposition of cell states\textsuperscript{148}, and reconstruction of simple one or two dimensional differentiation trajectories\textsuperscript{286,174}.

0.8.3 Research outline

The thesis is broadly partitioned into two parts: Chapters 1-3 focus on mapping fate choices in embryos, and between species. Chapters 4-5 apply similar methods to identify principles on how artificial systems differen-
tiate in culture. Two appendices provide supplementary data and a review of IncRNA mechanisms.

To develop the methods, we focused initially (Chapter 1) on the western claw-toed frog, Xenopus tropicalis, which is one of the best-studied model systems of early vertebrate development. After developing experimental conditions to harvest single cells from embryos with high viability and very little cell lysis, we collected a time series of 136,966 single cell transcriptomes spanning the first 24 hours of development post-fertilization. We then developed high-dimensional data analysis approaches to help annotate all of the cell types observed and to reconstruct developmental sequences. The resulting map showed an elaborate branching structure that revealed the waves of gene expression that gradually restrict the identity of each and every cell type in the embryo.

In Chapter 2, we then compare the differentiation map from frog embryos with a similar map from fish embryos. This comparison affords a unique opportunity to look deeply into these developmental gene expression programs and identify clues about the most consequential evolutionary changes across species. We identify topological rearrangements of several lineages and substantial gene expression divergence between orthologous cell types, and discuss the evolutionary implications.

In Chapter 3, we make use of access to both data sets to begin to ask systematic questions about patterns of cell fate choice in early development. We track the emergence of discrete expression states, show that fate choices often occur by refinement of overlapping gene expression, and show that transcription factors are reused pervasively across tissues.

In studying cell fate choice in cultured cells, we first (Chapter 4) use the experimental access of these systems to ask two main questions: i) are cell states sensitive to the method used to generate them? And ii) can the same cell state be reached by multiple distinct differentiation paths? We show that differentiation is path-independent by identifying two differentiation trajectories leading to the same terminal state.

In Chapter 5, we dissect the differentiation of a complex multi-tissue human organoid to: i) reveal the underlying differentiation pathways, and ii) identify mechanisms that can be manipulated to control differentiation outcomes. We show that organoid cell type composition is controlled by the dose-dependent activity of
the transcription factor GATA6 via an initial decision between differentiating into three alternative fates. We begin to leverage this mechanism to engineer specific mixtures of cell states.

We conclude by recapping the major findings in this work and considering several potential future directions.
A map, it is said, organizes wonder

Ellen Meloy

1

Mapping out embryonic cell states
1.1 Chapter overview

In this chapter we take the first steps towards scRNA-seq based mapping of differentiation in whole embryos. We focus on the Western claw-toed frog, Xenopus tropicalis, which along with its larger relative, Xenopus laevis, serves as one of the best-studied model systems of early vertebrate development. We first optimize experimental methods for the dissociation and transcriptomic profiling of whole developing frog embryos. We then apply these methods to profile embryos prior to the onset of zygotic transcription, through fate commitment and into the early tailbud stage - a point at which dozens of distinct cell types have formed encompassing the early progenitors of most major organs. In total we collect a time series of >130,000 cells spanning the first 24 hours of life post-fertilization. We curate this data set by annotating all of the cell types observed, and by developing computational approaches to reconstruct developmental sequences. We discuss in detail examples of new cell subtypes and new genes associated with fate choices. And we test how the neural crest emerges in early development, and resolve between two alternative models of how this lineage becomes pluripotent. Overall we develop general approaches that provide a global view of gene expression diversification in the early embryo.


Author contributions: The work presented in this chapter is my own work unless otherwise indicated. I worked in the lab to develop the novel dissociation conditions and to collect the scRNA-seq data. I worked on the computation analysis to establish the cell state classifications, map of cell state linkage over time, tests of neural crest formation models. I wrote the text and prepared the figures. The overall idea for the project was conceived jointly by myself and my advisors - Marc Kirschner and Allon Klein - when I first joined their labs. The project was made possible by Allon Klein’s post-doctoral work in the Kirschner lab; he invented the
droplet-microfluidic scRNA-seq approach (“InDrops”) used throughout this thesis. Several co-authors made important contributions. Marc Kirschner made the suggestion of varying pH to formulate the novel Xenopus dissociation buffer. Allon Klein and Caleb Weinreb provided suggestions for statistical and computational data analysis. Caleb also generated the web browser for viewing single cell data. Leon Peshkin provided support for frog experiments, and created the expanded gene symbol assignments used for our reference transcriptome. All authors provided valuable discussion and helped to edit the text.
1.2 Development of a dissociation media for Xenopus embryos

Xenopus blastomeres are large and fragile, with cell diameters up to 50μm (X. tropicalis, compared with 100μm for X. laevis) at the onset of zygotic transcription (stage 8.5). To preserve the integrity of cells through tissue dissociation and microfluidic handling, we optimized the inDrop platform to accommodate large cells. We also formulated a dissociation buffer that uses alkaline pH to promote dissociation, and handling procedures that minimize shear stresses placed on cells.

To identify the best possible dissociation conditions for Xenopus embryos prior to InDrops RNA sequencing, we initially performed a broad screen of known cell dissociation enzymes and buffers, including calcium magnesium free media (CMFM)\textsuperscript{209}, Newport dissociation media\textsuperscript{208}, trypsin-EDTA, TrypLE (ThermoFisher 12604-013), Accutase (Innovative Cell Technologies AT104), papain and collagenase. Five devitellinized embryos were incubated in each buffer at 18°C in 12 well plates, with gentle swirling by hand every minute initially, and then every 10 minutes until dissociation was complete or 1 hour had passed. All enzymatic dissociation reagents were ineffective on early Xenopus embryos. CMFM and Newport buffers led to substantial cell dissociation, but CMFM completely failed to dissociate the pigmented outer layer of cells. Newport buffer was deemed the best condition among those screened, but was still slow, requiring up to 1hr for near complete dissociation, and it frequently failed to completely dissociate small clusters of cells in each embryo.

Newport buffer had originally been formulated for the dissociation of pre-MBT Xenopus blastomeres. We attempted to optimize it further for post-MBT embryos. We systematically varied salt composition, pH, and buffering acid used. Newport buffer [0.1M sodium isethionate, 20mM sodium pyrophosphate, 20mM glucose, pH 9] could be made substantially more potent by raising the pH to 10.5 in the presence of 3-(Cyclohexylamino)-1-propanesulfonic acid (CAPS; a Good’s buffer with pKa=10.4) (Sigma C2632). In this new buffer – which we term “Newport 2.0” [0.1M sodium isethionate, 20mM sodium pyrophosphate, 10mM CAPS, pH 10.5] – embryos completely dissociated in <30min nutating at 80rpm and 18°C. Newport
2.0 effectively dissociated embryos from stage 8 (earlier stages not tried), through stage 22 (early organogenesis). After stage 22 certain tissues fail to dissociate completely, likely because of newly generated extracellular matrix. Dissociated cells were >95% viable for up to 2hrs at RT after washing with PBS-/- (Corning; 21-040-CV) as determined by staining with DAPI (wash protocol detailed below; DAPI added at 0.1μg/mL) (Supp. Fig. A.1). We also tested the extent of cell lysis by measuring free floating RNA in solution after dissociation by fluorometry with RiboGreen dye (ThermoFisher). After dissociation with Newport 2.0, free RNA was <5% of total embryo RNA, as judged by measurements of RNA in solution following dissociation compared to dilution series of RNA from whole embryos (Supp. Fig. A.1). We did not systematically investigate why high pH improves dissociation; it is possible that the high pH increases the dissociation constant of calcium ions bound to integrin receptors that hold embryonic cells together, or that it causes non-specific surface protein denaturation.

1.3 Single-cell RNA-seq of whole developing embryos

The samples for scRNA-seq were taken at ten time-points from zygotic genome activation (stage 8, 5hpf) through early organogenesis (stage 22, 22hpf), profiling a total of 136,966 single cells in two replicate experiments (see Table S1; Movies St-3 online). The first replicate consisted of 42k cells sequenced to a depth of 5.4k UMIs per cell on average (99% of genes from bulk RNA-seq observed in pools of >100 cells), whereas the second replicate of 95k cells was sequenced to an average depth of 1.4k UMIs per cell. The number of measured cells allows detection of rare sub-populations such as germ cells. Overall, we expect to observe transcriptional states as rare as 0.1% of the embryo represented by at least 10 cells with 95% confidence (Supp. Fig. A.2). There may be rare or transient sub-populations that are missed given the total number of cells, time sampling, and the depth of sequencing we used.

Low-dimensional visualization of the scRNA-seq timeseries shows a pattern of increasing complexity over time, with cells fragmenting from a continuum of gene expression states in the early gastrula into distinct
Fig. 1.1: Dissection of early Xenopus tropicalis development by scRNA-seq. scRNA-seq developmental timecourse including 136,966 single cell transcriptomes sampled over ten embryonic stages (S8-22). TSNE plots show increasing cell population structure over time. Colors indicate major tissues grouped by germ layer. Further details on subclustering shown in (Supp. Fig. A.3), and at: tinyurl.com/scXen2018.
clusters at later time-points (Fig. 1.1). To relate the scRNA-seq data to known embryonic cell types, we first classified gene expression states at each timepoint by a hierarchical clustering approach (Fig. 1.1B, Supp. Fig. A.3). These cluster assignments were robust to different clustering algorithms (Supp. Fig. A.4-5). We then annotated each cluster by matching cluster-specific genes to >2000 in situ measurements of marker genes for embryonic cell types documented on the Xenopus Bioinformatics database (Xenbase)\textsuperscript{132} (see Additional Data Table S1 online). We additionally shared our annotations with experts in the Xenopus research community, convened in part for that purpose at a Jamboree meeting in January 2018, who helped to standardize our cell state nomenclature and choice of annotations. In total, we identified 87 distinct cell type annotations, many of which persisted in clusters across multiple time-points (average of 2.9 timepoints each; Fig. 1.2C). Of these 87 states, 69 corresponded unambiguously to a specific Xenbase anatomical term, whereas the remaining 18 states correspond to finer structure (i.e. cell subtypes), or unidentified cell states within these anatomically defined tissues. Further substructure may be resolved by manual inspection of marker genes employing prior expert knowledge.

1.4 Reconstructing developmental cell state transitions

Fate mapping studies have previously established a set of spatial lineage relationships in the Xenopus embryo; these generally support the notion of a hierarchical pattern of differentiation, albeit with some exceptions\textsuperscript{57}. We investigated the extent to which these known hierarchical lineage relationships are reflected in our map of expression states. A close match was not assured since there are several reasons cell lineage hierarchies might not be reflected in molecular expression. First, cellular states are defined not only by tissue but by spatial position and cell cycle state, which cross lineage boundaries. Second, asymmetric cell divisions could lead to discontinuities in cell state as a mother cell partitions into daughter cells with distinct molecular compositions. Third, although the lineage history of individual cells consist entirely of discrete bifurcation (mitosis) events, cell states could be continuous or show complex branching structures or even loops. Yet,
counteracting this, in the lineage data of the early embryo, new gene expression generally arises in single contiguous domains.

To ask whether the single cell gene expression states conform to a branching pattern, and if so, whether the pattern overlaps with the previously deduced lineage patterns, we devised a simple algorithm that connects each cell to its most likely ancestors (i.e. nearest neighbors) in the previous time-point, and uses the consensus of these connections to assign an ancestor to each cell state (Fig. 1.2A). By applying this algorithm to all 259 cell states, comprised of 136,966 Xenopus cells spread over ten timepoints (Fig. 1.2B), we generated a detailed map of putative cell state transitions during early development (Fig. 1.2C, Supp. Fig. A.6). The map recapitulates known expression domains of master regulators within each germ layer over time (Supp. Fig. A.7). Significantly, only 17% of all votes fell outside of consensus ancestor clusters, showing that cell states over development can generally be approximated by a tree-like structure (Fig. 1.2D, Supp. Fig. A.6). The majority of votes cast outside of consensus ancestor states occurred at early timepoints, due to their continuum structure. The branching nature of cell states can also be appreciated by inspecting individual lineages at single cell resolution (Fig. 1.2E-F). The inferred cell state tree structure was consistent with ancestor assignments generated using an alternative state graph coarse-graining algorithm reported in our companion paper (Supp. Fig. A.8).

The ancestor assignments largely agree with the known lineage relationships. Of all cross-time-point edges inferred by the algorithm (Fig. 1.2C), we could confirm 234 (91%) by comparison to Xenbase anatomy ontology (XAO) and the literature (see Additional Data Table S2 online). Of the remainder, 22 could not be ruled in or out, and only 2 (1%) were identifiably incorrect. The two errors, which occurred in somites and erythroid lineages, reflected a specific limitation of the tree-building approach: in both lineages, mature states initially arise from progenitors, but then form a parallel branch, rather than continuously arising from progenitors anew at each time point. Visualization of the raw scRNA-Seq data reveals the underlying asynchronous differentiation process (Supp. Fig. A.9). In addition to these errors, the cell state tree does not consistently resolve spatial localizations, e.g. we do not resolve individual somites. The tree-building ap-
Fig. 1.2: Inference of developmental cell state transitions from gene expression similarity. (A) Schematic of the mapping algorithm used to make similarity connections between clusters across time. (B) Global visualization of single cells profiled in the Xenopus developmental timecourse using a kNN-graph (Wagner et al.). (C) Cell state tree showing all inferred developmental transitions. Generated by applying the mapping algorithm in A. (D) Representative cell voting outcomes between time points, generated during state tree construction. (E) Single-cell visualization of a representative sub-tree, showing lateral and intermediate mesoderm fates. Lines indicate corresponding topology of the cell state tree. (F) Marker gene expression associated with the formation of each intermediate mesoderm cell state.
proach could additionally generate errors when applied to other data sets lacking the dense time-sampling carried out here. When inferring cell state transitions from scRNA-seq snapshots separated by large gaps in time, intermediate progenitor states may be overlooked. Overall these results support a growing body of single-cell bioinformatics methods that seek to infer developmental cell trajectories on the basis of continuity and distance minimization in gene expression space \(^{286,309}\), but also illuminate where this principle can be misleading.

Among the developmental relationships inferred on the tree, there is also a question about the timing at which known cell types first appear. Of the 69 cell types with an unambiguous match to Xenbase, 60 appeared in our data at the developmental stage indicated in XAO \(^{247,248}\) or earlier. Several appeared much earlier than previously recognized (Fig. 1.3A), including an endothelial / hemangioblast progenitor, which appeared from the dorsal lateral plate region at Stage 18, as compared to Stage 26 / 31 (respectively) as previously thought \(^{167,302}\) (Fig. 1.3B). This afforded an opportunity to explore the earliest transcriptional events associated with the specification of these fates (Fig. 1.3C). Notably, the tail bud and multiple epidermal cell types also appeared much earlier than indicated by XAO (Fig. 1.3A, Additional Data Table S2 online) and revealed previously unappreciated early transcriptional dynamics.

Overall, the cell state tree provides a resource for probing gene expression in the early Xenopus embryo, incorporating annotations of cell states, linking related states across time, and discovering the earliest transcriptional record of the specification of each cell type. It is now possible to identify genes that are differentially expressed between bifurcating cell states, to assess gene expression changes along the developmental history of each lineage, and to automatically identify specific marker genes for each cell state (Fig. 1.4).

We have tabulated genes differentially expressed at every embryonic cell fate choice (see Additional Data Table S3 online), indicating potential fate regulators, as well as marker genes for every cell state (see Additional Data Table S4 online). The complete data set, which is available through an interactive online browser (tinyurl.com/scXen2018) supports visualization of gene expression across the cell state tree, identification of enriched genes in specific states, differential expression analysis between cell states, co-expression analysis
Fig. 1.3: scRNA-seq detects early transcriptional events during specification of embryonic cell states. (A) Time of first appearance for each cell state in the cell state tree as compared to documented appearance times in the Xenopus anatomy ontology (XAO). Red/blue points are detected early/late in scRNA-seq as compared to XAO. 60 of 69 states appear as early or earlier than documented. Error bars represent time interval of scRNA-seq experiment. (B,C) scRNA-seq reveals an early endothelial/hemangioblast progenitor that appears at stage 18 (red lineage), as compared to stage 26 for hemangioblasts and stage 31 for endothelial cells (XAO; [12, 13]), with recognizable activation of the endothelial/hemangioblast gene expression program (C).

of gene pairs, and visualization of the dynamics of gene expression along particular cell state tree lineages. Interactive plots of single cells from each stage are also available through this browser.

1.5 Assessing pluripotency retention during neural crest development

Single cell gene expression can help us to clarify specific developmental relationships. Among embryonic lineages, the neural crest is unique in its broad fate potential, contributing to multiple tissues spanning germ layer boundaries. It remains an open question how this unique fate potential arises. As the neural crest forms (S13/S14), it expresses at least eight pluripotency genes – foxd3, c-myc, id3, tfap2, ventx2, etsu, snai1, and oct25 – which are also expressed in the early blastula (S8/S9)18. Functional assays showed that several of these genes are required for multipotency of both blastula and neural crest cells18. This raised the intriguing possibility that neural crest may retain multipotency from the blastula stage (a “retention model”), by contrast to the classical view that a field of ectodermal tissue re-acquires multipotency during early neurulation (Fig. 1.5A). However, this model remains a hypothesis because a multipotent intermediate state between blastula and
Fig. 1.4: The Xenopus cell state tree is a powerful gene expression resource. (A) Global visualization of differential gene expression detected at every cell fate choice, showing the number of cell state specific genes (>4-fold enriched) relative to sister fates at each split. Tens to hundreds of fate-correlated genes are detected at every split. Panels B-D provide a more granular view of some of the gene expression information encoded within the intermediate mesoderm subtree (boxed). (B) Heatmap showing differentially expressed genes at each cell fate choice in the intermediate mesoderm. (C) Visualization of single cells (replicate 1) in the intermediate mesoderm subtree, showcasing examples of lineage specific TFs (also shown in Fig. 1.2E,F). (D) Reconstruction of the gene expression dynamics occurring during pronephric mesenchyme development. Cascades of early, middle, and late gene expression can be detected. (E) scRNA-seq data allows interrogation of gene coexpression in single cells, revealing correlated and anti-correlated gene pairs during fate choices. Pm, Cd as defined in grey box (panel B).
neural crest stages has not been directly identified.

Our single cell data offers an opportunity to re-examine these alternative hypotheses with a high-resolution view of the intermediate states between blastula- and neurula-stages in neural crest development. To support the “retention model”, we specifically searched for a sub-set of cells that maintain a blastula gene expression program, distinct from the remaining early ectoderm. To further increase the likelihood of finding such cells, we supplemented our whole embryo time series with additional scRNA-seq data (9,308 cells) collected from tissue that we dissected from the neural plate border region – that is fated to become neural crest – of stage 11 embryos, immediately prior to the expression of neural crest genes (total of 15,426 stage 11 cells). These data failed to reveal a distinct cluster of cells defined by expression of the proposed pluripotency genes, nor by any other set of genes enriched in the blastula. Instead, the data supported a more conventional differentiation pathway through clearly identifiable neuroectodermal intermediates (Fig. 1.5B). The same results held when examining differentiation progression not at the cell cluster level, but at single cell resolution (Fig. 1.5C). Far from showing a gene expression program that persists from blastula stage, we found that the inferred precursors of the neural crest showed large changes in gene expression with hundreds of dynamically varying genes (Fig. 1.5D). The proposed suite of eight pluripotency genes, in particular, were not limited to a single cluster of cells but were broadly expressed across most of the ectoderm, as well as in non-pluripotent states such as the endoderm and mesoderm at stage 11 (Fig. 1.5E; see Supp. Fig. A.10 for individual genes). This broad expression pattern would have been more difficult to appreciate without single cell data. The original experiments\textsuperscript{8} could not weigh equally the expression of across all tissues because the whole mount in situ staining used emphasizes superficial tissues.

We conclude that, at least at the level of transcription, there is no evidence of a distinct expression program that persists from the blastula to give rise to the neural crest. We fail to see such cells both using unsupervised approaches (clustering) and by examining the set of eight shared genes previously proposed to maintain pluripotency. It is not possible to completely rule out the retention of a blastula-like pluripotency program in neural crest precursors from our data: if functional pluripotency was maintained by chromatin state or
Fig. 1.5: Assessing the retention of pluripotency during neural crest development. (A) Contrasting models of neural crest development. Model 1: neural crest emerges from an intermediate population that retains blastula pluripotency (29). Model 2: neural crest emerges from ectoderm and reactivates pluripotency. (B) Ancestors inferred from scRNA-seq support model 2, where neural crest derives from neural cells at the neural plate border. (C) Single cell visualization (SPRING) of neuroectoderm, non-neural ectoderm, and neural crest also indicates that neural crest derives from the neural plate border. (D) Neural crest differentiation involves hundreds of >3-fold dynamic marker genes. (E) At stage 11, the shared pluripotency circuit proposed by Buitrago-Delgado et al. (29) - foxd3, c-myc (myca), id3, tfap2a, ventx2.1, ets1, and snai1 and pou3f5.2 - is expressed broadly in non-pluripotent cells. Score shows normalized aggregate expression; see Supp. Fig. A.10 for individual genes.
posttranslational modification it may be undetectable by scRNA-seq. Nevertheless, these new studies argue against a persistent transcriptional program preserved from an earlier stage uniquely in the neural crest. Rather, we argue in favor of a more conventional view of neural crest development proceeding from a well-defined ectodermal lineage.

1.6 Discussion

Embryonic development involves a carefully timed set of changes in cell behavior that drives the egg to a complex spatial and compositional pattern of cell types. A biochemical dissection of the underlying processes in development is complicated by limited material in the embryo and the heterogeneity of its composition, both of which must be considered together. Methods such as in situ hybridization have bridged these difficulties and can now be performed quantitatively and simultaneously over many genes. However, registering such spatial information at single cell resolution can be challenging, particularly in three-dimensional tissues. In an alternative approach, we and others have developed single cell transcriptomic methods, which sacrifice spatial information and in return provide a universal modality of measurement that is easily adapted to diverse situations. In a short time, these methods have been widely deployed and excelled in revealing cell population structures and dynamics. The resulting cell atlases can be computationally related to spatial structure using extensive in situ expression databases such as Xenbase. For embryos, applying scRNA-Seq involves important considerations. The method should not preferentially select a single cell type. It should be highly efficient in yield of cells. The dissociation procedure adopted should be quick and complete. Background RNA released by lysed cells in the sample should be minimized. These problems are acute for Xenopus embryos, as it has large yolk-filled blastomeres that are sensitive to shear forces and handling. Yet these problems have been overcome with a method of efficient capture and very little cell lysis.

We also developed analytical methods for interpreting the time course measurements of single cell gene expression in the embryo. Large temporal differences obscured temporal mappings between cell states using
established tools. This challenge was solved in our analysis by using shared latent spaces to measure similarities between states across adjacent timepoints. Upon linking cell states over time based on their gene expression, we found a strong match to known lineages. There were some differences that could be explained by the heterochronic formation of the same cell type, such as in the differentiation of the somite tissue, and in some cases the transcriptional cell states did not cluster along spatial boundaries. Yet in other aspects the data showed high sensitivity and could detect the first appearance of numerous cell states far earlier than previously appreciated. As shown for the neural crest, the sensitivity and single cell resolution of our gene expression resource can be used to test specific hypotheses about fundamental developmental processes.

We have made our annotated dataset a publicly available resource and are in the process of integrating our web browsers into Xenbase, the frog communities central database, where it will serve as a standard tool for researchers to query the cell type expression profile of any gene.

The approaches and analyses developed here open the possibility of a data-driven dissection of developmental programs and how they change across species. These two topics are explored further in chapters 2 and 3.

1.7 Materials and methods

Collection of Xenopus embryos

Xenopus tropicalis embryos were collected as described in\(^\text{309}\): mating pairs were injected with 200U of HCG per female and 50U HCG per male and allowed to mate naturally overnight at 22\(^\circ\)C. Embryos were dejellied in 2%wt/vol cysteine the next morning, 6hrs after injection, and sorted into batches of staged clutches pre-mid-blastula-transition (pre-MBT) and allowed to continue developing at 25\(^\circ\)C in 1X MMR. As the embryos reached the desired developmental stage, 5-10 embryos at a time were sampled from the clutch for dissociation and processing by InDrops. A single healthy clutch was used for the first InDrops time series; a further five independent clutches were used for the timepoints comprising the second replicate experiment.
Dissociation of Xenopus embryos into single cells for InDrops scRNA-seq

Removal of vitelline membranes: Vitelline membranes were removed by treatment with 1mg/mL pronase (Calbiochem) in 1X MMR, nutating at 80rpm and 18C for 8-10min. Forceps and inspection on a dissecting microscope were used to judge the earliest point at which the vitelline membrane had dissolved to minimize damage to the superficial layer cells of the embryos. Embryos were washed 2-3 times gently in 1X MMR immediately following pronase treatment to remove any vitelline membrane fragments stuck to the surface of the embryo, since carry over vitelline membrane interfered with dissociation.

Dissociation: A 12 well plate was pre-treated with Bovine Serum Albumin (BSA), and then a single well was filled with 5 mL of Newport 2.0 dissociation buffer [0.1M sodium isethionate, 20mM sodium pyrophosphate, 10mM CAPS, pH to 10.5 using NaOH]. Washed, devitellinized embryos were then transferred into the well. BSA pre-treatment was performed by air drying a film of 1mg/mL UltraPure BSA (Thermo Fisher) in ddH2o under a heat lamp. This treatment prevents dissociated cells from sticking to the tissue culture plastic surface. The 12 well plate well was filled to the top with Newport 2.0 buffer, and sealed using a layer of parafilm and silicone well isolator (Sigma; So810) after excluding all air bubbles. Sealing from air was essential to prevent cell lysis at the air-liquid interface during mechanical agitation. To mechanically accelerate dissociation, the sealed well plate containing fresh embryos was taped shut and attached to a benchtop Vortex genie using a plate adaptor. Embryos were agitated on speed setting 1-3 for 5 min, then accelerated to speed setting 7 for an additional 5-10 min or until dissociation was complete (typically <20 min). Embryos were entirely dissociated into monodisperse cells following this procedure. As soon as dissociation was complete, the plate was cooled on ice and single cells were allowed to settle to the bottom of the well by gravity for 5-10 min.

Washes and cell collection: Once settled, the single cell suspensions were washed by letting the cells settle by gravity through fresh buffer underlying the cells. Cells were first washed through a cold solution of 5%v/v Optiprep (Sigma; D1556) in 1X PBS-/ (Corning; 21-040-CV), then increasing to 10% v/v and 20% v/v. The buffers are added slowly against the bottom of the well, taking care to avoid cells and without agitating the
cells so that they would not rise to the air-liquid interface. Each increasingly dense layer of Optiprep replaced the bottom liquid around the cells, allowing complete buffer exchange without the need for centrifugation. Cells require 1-2 minutes to settle after each wash. After the washes, 1mL of 30% v/v optiprep with 1mg/mL UltraPure BSA (Thermo Fisher) in PBS/- was added as a loading solution for the cells. 30-35% v/v Optiprep is approximately neutral density for Xenopus cells, preventing them from settling during an InDrops run. To load the cells into a syringe for the inDrop platform, cells were floated off the bottom surface of the well by adding 0.5mL 40% Optiprep (denser than the cells) in PBS/-, causing them to accumulate at the interface between the 30% and 40% Optiprep phases. The cell layer was then aspirated directly into a 1mL syringe, which was backfilled with 0.1mL loading solution to avoid contact between cells and air, minimize cell handling, and block the plastic syringe surface. A wide bore needle tip and tubing (0.86mm internal diameter; SCI; BB31695-PE/5), pre-coated with BSA, was then attached to the syringe and connected to the InDrops microfluidic setup for cell injection. InDrops droplet collection was initiated immediately. The viability of dissociated cells remains >95% over 2hrs in loading solution.

InDrops single cell barcoding, library preparation, and sequencing

Dissociated Xenopus cells were processed by InDrops as described by Klein et al with the following changes: we minimized the distance between the cell injection syringe and the microfluidic chip; used a wider bore size tubing; and coated both the tubing and syringe with BSA prior to performing the experiment. These modifications minimize shear stress during microfluidic processing which can badly damage the large and fragile Xenopus blastomeres. After reverse transcription in droplets, barcoded emulsions were split into batches estimated to contain 2,500 cells, and chemically broken. Libraries were prepared as described previously. For the initial replicate of 42,385 cells, single cell libraries were pooled and sequenced together across seven NextSeq high runs, generating a total of 3 billion raw paired-end reads (see Table S1 online). The second replicate of a further 94,131 cells was sequenced across a total of ten NextSeq high runs achieving approximately half the sequencing depth per cell of replicate 1 (see Table S1 online). Within each repli-
cate pool concentrations were adjusted to ensure similar depth across all samples. Reads were demulti-
plexed using an updated version of the custom bioinformatics pipeline described in Klein et al\cite{143} available
at https://github.com/indrops/indrops, using a reference Xenopus transcriptome (XTr 9.o). The final output is an integer counts matrix of cells vs. genes.

**Expansion of gene symbol assignments in the XTr 9.o reference transcriptome**

Over 15k genes in the XTr 9.o reference transcriptome lack gene symbols, though in many cases an unam-
biguous protein identity can be identified by sequence homology. To fill holes in the gene symbol assign-
ments we assigned protein gene symbols to each XTr 9.o transcript using a modified reciprocal best HMM-
MER hit approach\cite{244} based on a target reference set of curated human proteins. Specifically, as a reference
set we used 25,208 Human and 4,492 Xenopus tropicalis, and 6,430 Xenopus laevis curated sequences ob-
tained from UniProt\cite{298}. Multiple matches with identical E value were resolved based on bit score. Addi-
tionally, sequences which did not fall into any bidirectional best match were assigned a symbol based on the
unidirectional best HMMER match. Only matches with an E value better than 1e-5 were considered. The re-
sult of our pipeline matched a total of 22,599 Xenopus tropicalis genes to one of 16,016 unique gene symbols,
which provides 11,341 more Xenopus genes with a gene symbol than available in the original genome annota-
tion, significantly enhancing data usability. We provide the enriched gene symbol assignments alongside the
XTr 9.o gene symbol assignments on our web-browser and in Additional Data Table S7 online.

**Data clean up: minimum expression threshold, doublet and background removal**

To remove background signal from putative empty droplets, only cells with >500-1000 detected UMIs, de-
pending on the library sequencing depth (see Table S1 online), were carried forward for analysis. After plot-
ting cells in two dimensions using tSNE (details below), we observed some clusters that appeared to consist
of cell doublets and others that likely represented empty droplets. We classified doublet clusters using three
criteria: i) lack of specific marker genes; ii) mixing of marker genes from other clusters; iii) a small size matching the experimental expectation of <2.5% double cell encapsulation during inDrops. Cells associated with doublet clusters were excluded. Background clusters were distinguished by i) lack of specific marker genes; ii) weak and uniform expression of all genes observed in the time-point. We noted that these background clusters form a characteristic ball-like shape in tSNE. All analyses presented in the paper represent cells passing these three filters (UMI>minCounts, not-doublet, not-background). We provide unfiltered gene expression data online and indicate which cells were excluded.

Visualization of data using PCA-tSNE and SPRING

To visualize high-dimensional single cell data in Fig. 1.1 we implemented the PCA-tSNE pipeline described in Klein et al.\(^\text{141}\), which follows these steps: i) normalize gene expression by the total transcript count per cell; ii) identify principally variable (PV) genes and non-trivial principal components compared to randomized data; iii) z-score normalize PV genes, perform principal component analysis (PCA) and retain the non-trivial principal components; (iv) run tSNE\(^\text{102}\) to generate a two-dimensional visualization. In addition to tSNE, we used SPRING as a complementary method to visualize single cell data (e.g. in Supp. Fig. A.5), as described in\(^\text{308}\). The tSNE perplexity parameter was set to 25 in all plots. SPRING builds a k-nearest neighbor (knn) graph from the high dimensional distances between cells and then embeds the graph in two dimensions using a force-directed layout, which emphasizes the continuous relationships between different cell clusters. For SPRING we used Euclidean distances computed from the same PCA coordinates used as inputs to tSNE.

Clustering of cell states

We first clustered cells from each time-point using local-density clustering (DBSCAN, as used in\(^\text{173}\)) applied to tSNE coordinates. We then used two steps to refine our clustering. First, we performed an initial round of clustering on all cells that over fragmented the data and then recombined adjacent clusters that had less than
3 genes expressed at >0.5 UMIs/cell on average that were >4-fold enriched in one state versus the other. Second, we isolated every cluster and performed a second round of dimensionality reduction and clustering on the cells for each cluster in isolation. The clustering approach can be summarized as follows: i) apply PCA-tSNE to the desired normalized counts matrix (see above); ii) use DBSCAN to generate an over-fragmented clustering; iii) manually merge adjacent clusters with insufficient differential gene expression – we refer to the clusters generated in steps (i-iii) as “level 1” clusters; iv) repeat steps (i-iii) for each level 1 cluster to generate sub-clusters, which we refer to as “level 2”. This process was performed on replicate 1 cells from stages 8 – 22 (all even stages), and on new timepoints from replicate 2 – stage 11 and 13. Other replicate 2 cells were clustered by a method that was informed by the initial clustering (see section below “Addition of replicate cells to cell state tree scaffold”). Three rare cell states were classified by manual inspection of marker genes as they were too few in number to form a distinct cluster recognizable by tSNE-DBSCAN: germ cells, neuroendocrine cells past St4, and hatching gland cells past St6.

Benchmarking against an alternative clustering method

We tested the robustness of tSNE-DBSCAN clustering by comparing clusters of stage 22 (replicate 1) cells to those generated by an alternative method based on spectral clustering with a kNN-graph kernel. We used the graph generated by SPRING, described above. Notably, the spectral clustering approach depends directly on the high dimensional distances between cells as opposed to their 2D tSNE coordinates. We found good agreement between the spectral clusters and those generated through our semi-manual tSNE-DBSCAN approach (Fig. A.4). Using an equal number of spectral and tSNE-DBSCAN clusters, a median of 79% cells in each tSNE-DBSCAN cluster mapped to its most similar spectral cluster, while a median of 89% cells in each spectral cluster mapped to its most similar tSNE-DBSCAN cluster. Most of the disagreements involved a single tSNE-DBSCAN cluster mapping to multiple spectral clusters (or vice versa), rather than a violation of cluster boundaries.

Annotation of cell state clusters
To relate the scRNA-seq datasets to known embryonic cell types we initially manually matched each gene expression cluster to Xenopus Anatomy Ontology (XAO)247,248 terms based on top marker genes. The XAO includes thousands of curated in situ staining measurements of tissue specific gene expression. By identifying genes that were specific to each gene expression state, and that had clear tissue specific expression patterns on XAO, it was possible to associate each gene expression state to a particular cell type (also see main text).

We then validated our annotations using a systematic approach as follows. (1) For each XAO term, we bioinformatically curated a set of literature-validated marker genes by parsing the gene expression database for Xenopus laevis and Xenopus tropicalis on Xenbase, found at the following locations:

ftp.xenbase.org/pub/GenePageReports/GeneExpression_tropicalis.txt
ftp.xenbase.org/pub/GenePageReports/GeneExpression_laevis.txt

We filtered for genes measured by RNA in situ hybridization only. (2) For each annotated single cell cluster, we identified marker genes as described below in the section “Identification of marker genes for each cell state”. (3) For each XAO term, we intersected the marker gene list from (1) with the cluster-specific gene list from (2). From these automatic marker gene annotations, all but nine states had at least one Xenbase-shared marker gene, of which five (dorsal lateral plate region, intermediate mesoderm – ssg1, trigeminal and profoundal placodes, small secretory cells, and epidermal progenitor – tp63/tll2) had matched anatomy terms with marker genes documented on Xenbase in the first place. 54% of annotations had at least three marker genes (compared to 5% when tissue annotations were randomized). The resulting list of literature-supported marker genes for each cluster is provided in Additional Data Table S1 online, along with literature references.

Connecting cell states across time from gene expression similarity

We investigated the developmental transitions connecting clusters across time points by asking each cluster to ‘vote’ on its most likely ancestor cluster from the previous time-point, using the following steps.

1. For each two adjacent time points, we embedded all cells from the two time-points in the PCA space learned from the second time point only, keeping non-trivial PCs as defined above. This embedding
causes cell-cell distances to reflect gene expression variation between tissues, as opposed to global changes over time.

2. In this embedding, for each cluster in the late time point, each constituent cell reported on the cluster identity of its 5 nearest neighbors from the previous time point using a Euclidean distance metric.

3. The number of edges to each early time point cluster were aggregated across all cells in each cluster. Percentages of votes cast for each possible ancestor are shown in heat maps on Supp. Fig. A.6, and the cell state tree presented in the main text represents vote winners. The winner of each vote was usually unambiguous, with <5% of assignments sharing more than 30% of votes with non-ancestor states, and a median vote-winning share of 88%.

We performed the ancestor voting process first for the “level 1” clusters (see section on clustering for a definition) and then, for level 1 clusters that had been sub-divided into level 2 clusters, we performed a second – private – voting process on the level 2 clusters – i.e. between only subpopulations contained within connected level 1 states. This two-step procedure was necessary because the level 2 clusters are only distinguishable with PCA coordinates learned from each level 1 cluster separately, and therefore had to be bioinformatically isolated when computing inter-time point cell-cell distances.

**kNN-graph visualization of entire scRNA-seq timecourses**

Whole-embryo kNN-graphs were generated using the approach reported in Wagner at al. In Supp. Fig. A.8 we compare the Xenopus and Zebrafish cell state trees to this alternative representation.

**Addition of replicate cells to cell state tree scaffold**

Late during the preparation of this manuscript, we performed a large replicate experiment, that added a total of 94,131 additional single cell transcriptomes to the original time series. The replicate included all original timepoints except stage 10. It also included two new intermediate timepoints – S11 and S13. All cells were
quality filtered by the same process as the original experiment. New timepoints (S11 and S13 only) were clustered and mapped into the cell state tree structure by the same methods as the original experiment. New data from existing timepoints (all but S11 and S13) were clustered by a distinct process that (i) first assigned every cell to a pre-existing annotated cluster; (ii) performed a batch correction of gene expression between the biological replicate experiments; and then (iii) examined the combined data set for novel sub-structure within the annotated clusters. The final annotations presented in this paper incorporate new sub-structure found through this process. The detailed steps are as follows:

1. **Assign cells to a pre-existing annotated cluster:**
   
   (a) New data at each time point was fragmented into \(4^N\) clusters, where \(N\) is the number of clusters in original experiment at the matched time point. Clustering was performed by k-means spectral clustering on the kNN graph, constructed using parameters described above (see SPRING visualization).

   (b) Each cell in the new data was assigned a cluster label based on the original cluster annotations, by connecting the \(4^N\) new clusters to the original \(N\) clusters within each time point. The assignment algorithm is as described for cross-timepoint mapping (see previous Methods section; median voting consensus of 80%).

2. **Batch correct biological replicates:** This method is a variation on the batch correction algorithm recently proposed by Marioni et al. in pre-print (https://doi.org/10.1101/165118). From step (i), each cell in both biological replicates is assigned to a unique cluster \(i\). We defined the gene expression vector for cell \(j\) in cluster \(i\) from replicate 1 as \(x_{ij}\), and respectively from replicate 2 as \(y_{ij}\). The centroids of cluster \(i\) in the two replicates (average of all cells in the cluster) \(x_i, y_i\) respectively, were calculated in units of TPMs from raw counts. We further defined the arithmetic mean of both centroids \(z_i = (x_i + y_i) / 2\), and a multiplicative correction factor \(\Delta_i = (1 + x_i) / (1 + z_i) - 1\). The corrected gene expression values for each cell \(j\) in cluster \(i\) were then calculated as \(x'_{ij} = (1 + x_{ij}) / (1 + \Delta_i) - 1\), and
\[ y_{ij}' = \frac{(1 + y_{ij})}{(1 - \Delta_i)} - 1. \] Corrected counts matrices \(x\) and \(y\) from both experiments were combined and carried forward for further analysis. This correction procedure centers the centroids of cells in each cluster from the two biological replicates through a multiplicative correction, ensuring non-negative and non-infinite values for gene expression. After correction, both replicates intermixed completely.

3. **Discover substructure within pre-annotated clusters:** each new cluster, now composed of cells from both replicates, was sub-clustered in isolation by tSNE-DBSCAN (see above). If only a single cluster was identified, then no sub-structure was added to the annotation. If more than one cluster was identified, the new sub-clusters were examined for expression of marker genes of the original annotated cluster. If a sub-cluster lacks typical marker genes of the tissue, the cluster was excluded as outlier cells. All sub-clusters expressing marker genes of the original annotated cluster were assigned novel annotations and included in the final state tree.

**Benchmarking of cell state tree similarity relationships against XAO**

We manually inspected each of the 257 edges in the resulting cell state tree, using the XAO database and literature to determine the validity of each edge. Each edge in the tree links a parent cluster in time point \(t\) to a daughter cluster in time point \(t+1\). We tested three criteria: (a) the daughter cell state is believed to arise from the parent cell state; and (b) if the daughter cell state differs in XAO annotation from the parent, then the daughter state should appear at the time specified in the XAO database indicated in the Start_Stage field; (c) if the daughter cell state differs in XAO annotation from the parent, then the parent state should end at the time specified in the XAO database indicated in the End_Stage field. For each edge, we provided at least one reference to XAO or to the literature supporting our verdict. We assigned each edge one of seven flags: “OK” indicating that all criteria are satisfied (203 edges); “Discovery: early” indicating that a state appears prior to its annotated first appearance in the XAO (118 edges); “Starts too late” indicating that a state only appears after its annotated first appearance in the XAO (9 edges); “Ends too early” indicating that the parent state
terminates prematurely (1 edges); “Ends too late” indicating that the parent state terminates after expected in the XAO (3 edges); and “Error” indicating that the daughter does not arise from the parent state (2 edges).
In addition, 21 edges were assigned “N/A” as they related to the appearance of daughter states that did not correspond to XAO terms. The full table is provided in Additional Data Table S2 online and the difference in timing of appearance is plotted in Fig. 1.3A.

Identification of marker genes for each cell state

The 259 clusters in the data were mapped to 87 unique annotations in the Xenopus cell state tree, many appearing over multiple time points (also see above section “Annotation of cell state clusters”, and main text). We defined marker genes for each of the 87 unique annotations in the Xenopus cell state tree with the following criteria: i) average expression enriched >5-fold compared to the average for the rest of embryo, ii) expression detected in >15% of cells in the state, and iii) the state has the highest average expression for the gene. We identified 2,159 marker genes across all states, with a median of 6 marker genes per state. The mean enrichment for the most specific marker gene for each state was 93-fold, and the mean expression level across all identified marker genes was 2.2 UMIs (930 TPM), a value approximately half the average expression level of the ‘housekeeping gene’ beta-actin (5.65 UMIs/cell or 2,360 TPM). Marker gene lists are tabulated in Additional Data Table S4 online.

Differential gene expression at cell fate choices

Differentially expressed genes between two cell states were identified using two-tailed Wilcoxon Rank Sum Tests, at a stringency of p<0.001 with multiple hypothesis testing correction and FDR of 5%. Before rank sum tests were performed data was total count normalized across states, and filtered for genes that were expressed in >=10 cells, had an average of >=1 UMI/cell in at least one state, and were >=4-fold differentially expressed between states. Overall this procedure is highly conservative, requiring both high expression and high fold-change in expression; it identifies the most prominent differences between states efficiently and
with a low false positive rate. Differentially expressed genes at each fate choice are tabulated in Additional Data Table S3 online.
Nothing in biology makes sense except in the light of evolution.

Theodosius Dobzhansky
2.1 Chapter overview

There are numerous cladograms of gene sequence that confirm the basic paleontological record of vertebrates and reveal the genetic differences that separate species. Since the early 19th century, the similarities and divergences in embryonic and adult anatomy across species have also been extensively characterized. However, in between the genetic and anatomical scales, we have limited knowledge about how cell types and their developmental programs change across species. This gap limits our ability to understand how changes in genotypes lead to changes in phenotypes across evolution.

In this chapter we perform a comparison of two vertebrate species - frog (Xenopus tropicalis) and fish (Danio Rerio) - at the level of single cell transcriptomes. As a first step, we align the frog and fish cell state trees, highlighting similarities and differences in the topology of differentiation trajectories across species. We then use this map to ask: i) whether orthologous cell types have maintained their transcriptional states across 435 My of evolution; and ii) whether specific functional or sequence properties predict whether genes will show conserved expression patterns across species.


*Author contributions:* The analysis, figures, and writing in this chapter is my own work unless otherwise indicated. Caleb Weinreb helped to generate Fig. 2.2A and 2.5. The comparison relied on annotated scRNA-seq timeseries data from zebrafish that was kindly shared by Dan Wagner and Sean Megason prior to its publication. Leon Peshkin helped to identify gene orthologs between the frog and fish reference transcriptomes. Marc Kirschner and Allon Klein helped to edit the text.
Fig. 2.1: Frog and fish single cell maps. kNN graphs visualizing single cell data from frog (Briggs et al.; Science, 2018)⁸ and fish (Wagner et al.; Science, (2018)³⁰¹.

2.2 Alignment of frog and fish cell state trees

Given scRNA-seq timeseries data from the frog⁸ and the fish³⁰¹ (Fig. 2.1), we aimed to compare differentiation into each cell type across species. We processed the zebrafish time series through the same tree building algorithm applied to frog in the previous chapter. We manually aligned cell states between species (Fig. 2.2A; red shading) on the basis of tissue name, marker gene expression, developmental stage, and lineage relationships. We tested that no permutation of this matching improved the overall gene expression correlation between species, indicating high accuracy. In some cases, nomenclature of homologous tissues differed slightly between species. Thirty tissues could be matched with high confidence, jointly covering 66 of 87 Xenopus states (79% cells) and 83 of 122 zebrafish states (83% cells).

The cell state tree alignment indicated broad conservation of lineage topologies between species, although the observed abundances of cell types differed markedly between species, and the proportion of matched
Fig. 2.2: Alignment of frog and fish cell state trees. (A) Xenopus and Zebrafish cell state trees aligned by orthologous cell states (red shading) reveals similarities and differences in cell state hierarchies. Grey/white stripes provide a visual guide. (B) Single cell visualization of matched epidermal subtrees in frog and fish reveals in detail the similarities and differences in developmental hierarchy. The Xenopus and Zebrafish share conserved epidermal and ionocyte cell states, marked by conserved expression of gata2 and foxi1 respectively. By contrast, species specific states such as ciliated epidermal cells, small secretory cells (SSC), and goblet cells in Xenopus, are show species specific marker gene expression. SCC=small secretory cell; NE=neuroendocrine cell. Unidentified zebrafish cell types are labeled by marker genes on the cell state tree.
states and cells decreased over time as species-specific features accumulated – such as specialized epidermal cell types in both species (Supp. Fig. A.11). Notably, the zebrafish neural ectoderm grew to 60% of sequenced cells in the embryo at 24hpf, compared to 31% in the Stage 22 frog (Supp. Fig. A.12). Differences in cell survival rates during InDrops processing in either species could affect estimates of cell state abundances.

2.3 Changes in lineage topology across species

The alignment of the cell state trees also highlighted significant changes in developmental patterning between the two species, which could further be appreciated by deeper inspection of the underlying single cell data. In the epidermis, for example, the two species show a radiation of common and distinct cell types from gata2-expressing non-neural ectoderm (Fig. 2.2B). Trajectories of differentiating ionocytes can be seen in both species, whereas ciliated cells expressing foxj1, goblet cells expressing itln1, and small secretory cells expressing met (Fig. 2.2B), each of which is an important component of the Xenopus epidermis, are absent in zebrafish. These different epidermal cell types likely reflect the demands of development in different environments, potentially requiring an early immune barrier in the frog but not the fish. Rearrangements of lineage topology can also be appreciated: Xenopus and zebrafish both produce specialized hatching glands (HG) that secrete shared hatching enzymes and that are specified by the conserved transcription factor klf17, seen by scRNA-Seq (Fig. 2.3). Single-cell trajectories link the HG to non-neural ectoderm in Xenopus, but to the organizer mesoderm in zebrafish (Fig. 2.3A), reflecting a known difference in HG germ layer origin between Amphibians and Teoleosts.

Other more subtle changes in lineage topology involved the addition or loss of specialized progenitor cell states within a partially shared developmental history. For example, xanthoblasts appear at 18hpf in zebrafish from an early neural crest population, but do not appear until stage 46 in Xenopus, after the shared ancestral neural crest population has progressed through multiple additional specialized neural crest intermediate states. Similarly, myeloid cells appear in Xenopus at S14 from early involuted ventral mesoderm, whereas
Fig. 2.3: Topological rearrangement of hatching gland differentiation trajectory across species. (A) Aligned Xenopus and Zebrafish cell state trees indicating the hatching gland differentiation path in both species (red). The unsupervised tree building algorithm identifies that the hatching gland emerges from non-neural ectoderm in frog but mesodermal organizer tissue in fish. (B-C) Comparison of gene expression in the hatching glands of both species. scRNA-seq detects conserved expression of the TF, klf17, which is necessary for hatching gland differentiation in both species (B), as well as conserved expression of the hatching enzyme, he1a (C). Novel conserved and species specific gene expression is also detected in both species.
they (macrophages and leukocytes) appear only at 18-24 hpf in zebrafish, after the ventral mesoderm has differentiated through specialized cardiovascular and lateral plate intermediate states. Terminal differentiation of myeloid cells in both cases involves activation of several conserved master regulators including cybb, cyba, spib, and cebpa. These observations support a degree of independence between the differentiation path characterizing a lineage and the activation of particular conserved terminal cell phenotypes. Variation in differentiation paths can reflect not only drift or flexibility in developmental programs, but also selected features of embryonic cells including roles in interacting with and instructing other differentiation and morphogenetic mechanisms.

2.3.1 Changes in gene expression across orthologous cell states

We next asked whether gene expression is preserved in orthologous cell states between frog and fish (shaded pink groupings in Fig. 2.2A). Evo-devo has discovered numerous examples of deeply conserved gene expression across species, such as the conserved expression of Pax6 in the eyes of species as divergent as flies and humans. We therefore expected that high gene expression similarity may identify cell state orthologs between frog and fish. To test this, we computed pairwise gene expression correlations between cell states in frog and fish (Fig. 2.4; Supp. Fig. A.13). To our surprise, we found that gene expression similarity does not predict cell type orthology. While many states such as neurons, muscle, and non-neural ectoderm were significantly more correlated with their orthologs than with other states (Fig. 2.4; red blocks on the diagonal), many other cases cell states did not match to their orthologs on the basis of gene expression similarity. The frog ureteric bud, and ventral mesoderm, for example, were not correlated to the fish ureteric bud or ventral mesoderm (rho < 0.1), but were both significantly more correlated with the zebrafish neural crest (rho = 0.4). This was not due to noise in our comparison method, as correlations within a species after subsampling the data showed strong self matching between ortholog states (Fig. 2.4; A.13). These results therefore indicate significant divergence in gene expression in orthologous cell states over the 435 My of evolution separating frog and fish.
Fig. 2.4: Expression similarity does not predict cell type orthologs. Heatmaps show the Pearson correlation between matched cell states (replicate 1) for the indicated species-comparison and pre-filtering stringency. Genes that are predictive of cell state within one species tend to be poorly predictive across species. Self-correlation is computed by jackknifing the data within a species (see Materials and Methods).
To obtain a systematic view of similarities and differences in developmental gene expression across species, we next tested whether orthologous genes show correlation in their expression values across matched embryonic tissues. Focusing on the subset of genes that were robustly expressed and dynamically varying in both species, we observed that orthologous genes correlated in their expression between species (Fig. 2.5A), seemingly supporting the notion that cell state gene expression is largely conserved. However, only a minority (30%) of orthologous pairs were correlated with high confidence (5% FDR) compared to random pairs. For example, in contrast to the well conserved \( r=0.9 \) neuroectodermal marker sox2, we identified many examples of poorly correlated genes such as the transcription factor gata5 \( (r=0.1; \text{Fig. 2.5A}) \). Despite being a conserved endoderm and cardiac master regulator, gata5 is also expressed in the erythropoietic tissue only of Xenopus\(^{307} \), and in the marginal zone only of zebrafish\(^{29} \) at a distinct spatiotemporal location of endoderm specification compared to the vegetal expression seen in Xenopus. Accordingly, we were surprised to find that the genes best marking cell states within one species did not generally perform well at marking the same cell states within the other.

### 2.4 Properties of conserved genes

We tried to understand what properties predict whether a gene shows conserved expression patterns across tissues across species, i.e. whether specific functional categories of genes are more likely to be conserved in expression dynamics. Analysis of gene ontology annotations spanning diverse cellular processes revealed a striking enrichment of gene expression conservation among transcription factors, identifying three statistically significant \( (p<0.05; \text{binomial-test, Bonferroni corrected}) \) functional annotations that include “nucleus”, “regulation of transcription” and “transcription factor activity” (Fig. 2.5B). By contrast, protein sequence conservation was uncorrelated to gene expression conservation (Fig. 2.5C; \( r=0.01; p=0.6 \)). Orthologous genes with 20-40% sequence identity had the same expression conservation as those with 95-100% sequence identity (average expression conservation of 0.41 versus 0.38, \( p=0.5; t\)-test). A gene’s function is therefore more
Fig. 2.5: Properties of conserved genes. (A) Ortholog genes across species have variable conservation of cell state specific expression. Just 30% of self-similar orthologs are conserved at a 95% FDR compared to random gene pairs. Right panels: examples of highly (Sox2) and poorly (Gata5) correlated TFs across species. (B,C) Function, not sequence, predicts gene expression conservation: (B) Orthologs with highly conserved expression patterns across species are enriched in TF-associated GO terms. P-values show Bonferroni-corrected binomial test results. (C) Protein sequence conservation is not correlated with gene expression conservation (r=0.01; p=0.6).
strongly predictive of its conservation in developmental gene expression programs across species than conservation of its protein sequence, decoupling two fundamental processes underlying evolutionary changes in embryonic development.

2.5 Discussion

By comparing frog and fish single cell atlases, we identified surprisingly large divergence in cell-state-specific gene expression across species. Although the overall structure of development, including the formation of germ layers, and the patterning of each germ layer into subtypes, is broadly similar across species, we show that orthologous cell states between species in many cases could not be matched by gene expression similarity. We also identified instances of cell differentiation that were unique to one species, such as in the epidermis of the frog, and were able to comprehensively document the underlying molecular changes. These records provide a blueprint for how novel cell types may emerge in evolution. We also identified cases where the same cell state was produced by topologically distinct differentiation trajectories across species. This finding suggests that cell differentiation may be path-independent, and show flexibility in the developmental history that leads to each mature cell state. We will explore this idea further in Chapter 4.

Although the tissue-specific expression of individual genes was relatively variable across species, we did detect evidence of conservation of gene expression patterns across tissues. We found that orthologous cell states maintained expression of a subset of genes, which were most notably enriched for TFs, and among TFs, those that are used in a single tissue rather than those that are combinatorially reused across lineages. This finding supports previous arguments that cell identities may be maintained by a small number of genes, while phenotypes can vary substantially across evolution. We also found that gene expression conservation is independent of variation in protein sequence. This observation decouples a gene’s structure (and therefore function) from its expression pattern in the embryo across evolution.

The approaches and analyses presented here establish important first steps towards a data-driven dissec-
tion of how developmental programs change across species. As scRNA-seq data become available for more species, it will become possible to track the dynamics of gene expression divergence across evolution, and to pinpoint the precise changes in developmental programs that underlie the acquisition of novelties across species. These insights will help to bridge the gap between variation in genotypes and the resulting changes in phenotypes.

2.6 Materials and methods

Alignment of Xenopus and Zebrafish cell state trees

Xenopus and Zebrafish cell states were matched manually to enable computational comparisons of gene expression across species. Many-to-many matches of cell states were allowed to avoid ambiguities in matching timepoints across species, and to allow groups of states showing greater substructure in one species that the other to nonetheless be compared (e.g. the zebrafish spinal cord was divided into several subpopulations compared to the single Xenopus state). The criteria to match states were that they should: i) share a common position within the developmental hierarchy of both species according to Zfin and Xenbase anatomy ontologies, ii) share at least 3 validated marker genes, and iii) appear at approximately the same time across species. We sometimes encountered matchings where: 1) the name used by the frog community (Xenbase) differs slightly from that used by the zebrafish community (zfin); or 2) the time point at which we profiled the two species lead to one corresponding tissue having a less mature name as compared to the other (despite highly similar gene expression). To bioinformatically validate the manual matching we randomly permuted the cell state matchings and asked whether any changes could improve the mean gene expression correlation between species by >0.1, indicating a potential error. This revealed no errors in the manual matchings.

Gene expression correlation between species

Gene expression conservation between species was measured using spearman correlation of the average normalized UMI counts for orthologous across 33 orthologous tissues. Each orthologous tissue contains one or
more clusters from Xenopus and Zebrafish respectively, and the selection of clusters was performed manually with attention to marker gene expression, name, position in the embryo and inferred potential to produce downstream lineages. We created a list of orthologous genes and their percent sequence identities by selecting reciprocal best hits from a mutual amino acid BLAST search between all Zebrafish genes (assembly GRCz10) and all Xenopus genes (assembly XTr 9.0). Of these reciprocal best hits, only a subset consisted of genes that were robustly expressed in the single data. Reasoning that cross-species correlation of gene expression cannot be expected to exceed within-species correlation between different replicates, we computed a ‘self-correlation’ score for each gene by performing 50 random splits of the data from each species and recording the correlation of cluster-averaged expression for each gene in one half to itself in the other half. Only orthologous gene pairs where both genes had a self-correlation >0.75 were carried forward for analysis.

Analysis of functional gene categories enrich for conserved genes

Annotations of gene function, obtained from Gene Ontology (GO) term associations reported on Xenbase, were ranked by the proportion of constituent genes having a cross-species correlation of >0.66, with this threshold chosen to establish an FDR of 1% against a null distribution generated from randomly matched (i.e. non-orthologous) genes. We restricted the analysis to orthologous gene pairs having a ‘self-correlation’ greater than 0.75, as explained above, and focused on GO terms with at least 20 genes among this set. P-values were assuming a binomial null distribution for the number of conserved genes among those associated with each GO term.
The only simplicity to be trusted is the simplicity to be found on the far side of complexity.

Alfred North Whitehead

Conserved patterns of cell fate choice
3.1 Chapter Overview

The single cell gene expression landscapes generated in Chapters 1 and 2 provide new opportunities to examine gene expression changes associated with differentiation into dozens of cell types simultaneously, in single cells, and in the unperturbed setting of the embryo. To begin to explore these opportunities, in this chapter we ask three questions: 1) on what timescale do discrete gene expression states emerge in the early embryo; 2) how at the level of genes do new cell identities emerge at branch points in development; and 3) given the close association and conservation of transcription factors with differentiation, how are transcription factors deployed across cell types and across time during development? We leveraged data from both the frog and the fish to identify conserved fate choice dynamics.


Author contributions: The analyses, figures, and text in this chapter are my own work unless otherwise indicated. Sam Wolock helped to perform the analysis of discrete state formation in Fig. 3.1. Caleb Weinreb generated Fig. 3.3. In Fig. 3.2 and 3.4, we use zebrafish scRNA-seq data and annotation shared by Dan Wagner and Sean Megason prior to publication. Allon Klein and Marc Kirschner helped to edit the text.
3.2 Appearance of discrete gene expression states

Single cell gene expression data can be used to ask whether cells exist in a continua of gene expression states, or as discrete clusters.\textsuperscript{297,231,83} We used our timeseries data to track the dynamics of discrete state emergence during patterning of the embryo. We assigned each cell a discreteness score, based on the shape of its distance distribution to all other cells (Fig. 3.1A-B). The score is high for cells that are near to some cells, but very far from all other cells, indicating discreteness. By plotting the discreteness score of cells on a single cell graph of the whole embryo, we confirmed that the highest scoring cells were located in discrete projections of the graph, suggesting that the score identifies discrete expression neighborhoods.

At MBT, almost all cells had discreteness scores \(<1.5\). During gastrula and neurula stages, cells with discrete gene expression emerged rapidly (Fig. 3.1C), so that by stage 18, 45\% of cells had discreteness scores \(>1.5\) compared to \(<5\%\) of cells at stage 10. The discreteness score of cell states was widely distributed at later timepoints, with several highly discrete clusters (discreteness score \(>4\)), but also many continua clusters (discreteness \(<1.5\)). Epidermal and early myeloid states for example were highly discrete, matching their early maturation. The neural plate was still largely a single continua in gene expression by this metric even by stage 22, reflecting a more gradual mode of patterning and cell fate choice. We next asked how early continua might separate into discrete states at fate choices.

3.3 Multilineage gene expression priming during fate choices

It has been shown in several specific cell types – including hematopoietic stem cells\textsuperscript{114} and mouse ES cells\textsuperscript{284} – that uncommitted cells may coexpress competing fate regulators associated with more than one terminal cell fate. Such gene expression overlap, followed by refinement as cells differentiate, is known as multilineage priming (MLP) (Fig. 3.2A). MLP has been argued to be a part of the process of cell type decision making, via competition between fate regulators, and also to play a role in specifying an undifferentiated state by blocking lineage commitment\textsuperscript{156,217,284,297}. 
Fig. 3.1: Dynamics of discrete cell state appearance in Xenopus in scRNA-seq timecourse. (A) For each cell, a ‘discreteness score’ was calculated as the log-kurtosis of its distance distribution to all other cells in gene expression space (see Materials and Methods section “Cluster continuity analysis” for further details). The premise of the score is that high kurtosis can indicate bimodality, or discrete clustering of the distance distributions. (B) Discreteness scores plotted on the all-cell kNN-graph from Fig. 1.2B, showing that scores are high on relatively discrete cell state projections. (C) Dynamics of discrete cluster emergence. Grey dots indicate the average discreteness score of each cell state at each timepoint. Discrete states with scores > 2 (higher than the right tail of the stage 8 distribution) first appear at stage 12 and accumulate over time. Continua are also apparent at each timepoint. The large black dots indicate the fraction of total cells at each timepoint with discreteness scores > 1.5. Scores shown are calculated on replicate 1 data.
The significance of MLP for cell fate decisions is clearly supported by these specific examples, but its importance would be further strengthened if it were generally associated with branch points in differentiation. Our scRNA-seq data offered an opportunity to assess this over entire embryos. Wardle and Smith had previously\textsuperscript{106} noted transient overlap and then refinement between the ectoderm-specific TF Sox2, and the mesoderm-specific TF Brachyury (T) during gastrula stages in Xenopus. This was recapitulated in our data, which shows co-expression of these markers in 26% of cells expressing either gene at stage 10, but refinement to just 3% overlap by stage 14 (Fig. 3.2B,C). Our data also show numerous other examples, including similar MLP between early ectodermal TF Zic1, and the mesodermal TF Foxc1, which overlap in expression at stage 10 (7% coexpression) before refinement by stage 12 (<3% coexpression) (Fig. 3.2B). Across all fate choices in the embryo, we identify 412 MLP genes in Xenopus (see Additional Data Table S6 online). We find that MLP is initially widespread, encompassing a considerable fraction of highly-cell-state-specific genes at each branch point (>4-fold differentially expressed), but reduces over time (Fig. 3.2D,E). For fate choices that occur before or up until the end of gastrulation, over 70% of highly-cell-state-specific genes overlapped in their progenitor state in fish and 50% in frog. By contrast, for subsequent fate choices, MLP ultimately drops to <10% in both species (Fig. 3.2D). All differentiation branch points showed MLP before gastrulation, whereas only selected branch points showed frequent MLP after gastrulation (Fig. 3.2E).

Where MLP did occur, we wondered if it reflected generally noisy gene expression, or more specifically co-expression of regulators of cell fate. Suggestive of the latter possibility, MLP genes were significantly enriched for TF-associated GO terms including ‘transcription factor activity’, ‘regulation of transcription’, ‘nucleus’, and ‘DNA binding’ as compared to all DE genes (Fig. 3.3; 2-fold enriched; p<0.001 binomial-test, Bonferroni corrected).
Fig. 3.2: Refinement of promiscuous multilineage gene expression during early embryonic fate choices. (A) Illustration of multilineage priming (MLP). Two genes, specific to daughter states A and B respectively, transiently overlap in the ancestor progenitor state as a fate decision is being made. (B,C) MLP during the fate choice between neural plate and dorsal marginal zone in Xenopus. Sox2 and T, as well as Zic1 and Foxc1, overlap in progenitor cells before becoming specific. (D,E) Global patterns of MLP in early development: (D) Multi-lineage primed genes are initially pervasive among differentially expressed (DE) genes at fate branch points, but become progressively rarer. (E) MLP frequency shown for each cell fate choice on the cell state trees indicates sporadic MLP at later time points.
Fig. 3.3: MLP genes are enriched for TF-associated GO terms. MLP genes were compared to all DE genes (dark grey bar) across embryonic fate choices in Xenopus. Enriched TF-associated terms include 'Transcription factor activity', 'Regulation of transcription', 'Nucleus', and 'DNA binding'. P-values indicate significance as determined by a binomial test with Bonferroni correction.

3.4 Re-use of developmental transcription factors

Classical experiments have described differentiation in terms of the activation of single transcription factors, called master regulators, which established cell identity and after which other transcription factors progressively refined the phenotype [30] (Fig. 3.4A). Alternative views, however ascribe the initial specification of cell identity to interactions among multiple transcription factors, a ‘combinatorial code’, capable of defining a larger number of identities with different combinations from a small set of TFs [198,219] (Fig. 3.4B). We asked which of these views best describes early development. This is not just of theoretical interest since knowing which transcription factor(s) could specify a given cell type could aid in the formulation of combinations that could be used therapeutically [50]. Our experiments identify some cases where candidate TFs are activated just once in early development and after which they persist and may define a lineage, suggesting a master regulator model. But we also see TFs that are expressed more than once de novo in distinct and unrelated cell states, which may support a model of combinatorial deployment.
To understand the prevalence of these modes of TF use, we scored the number of times a TF was expressed independently, i.e. in states that share a non-expressing common ancestor. From this analysis, we found that the expression of half of developmentally-variable TFs is initiated only once during early development (52% in frog, 54% in zebrafish) (Fig. 3.4C). The detected single-use TF induction events occurred in just 35% of annotated cell states, seemingly inconsistent with the view that distinct master TFs define each fate. Reused-TFs, by contrast, cover 58% of tissue annotations in a complex combinatorial code (see Additional Data Table S5 online). In the remaining 7% of cell states, a specific TF induction was not detected. Over time, the average fraction of transcription factors being reused in a second circumstance climbs, reaching 90% of new TF expression initiation events by stage 20 (18.5 hpf) in Xenopus and by 20 hpf in zebrafish (Fig. 3.4D). Together these results underscore the importance of combinatorial TF reuse in developmental gene expression programs.

Consistent with the notion that combinatorial interactions could lead to complex gene expression responses, the expression of the same TF in different tissues did not generally correlate with the same gene sets, as seen in the example of Pax8, which is independently expressed in the otic placode and pronephric mesenchyme (Fig. 3.4E). Some putative downstream targets of re-used TFs could however be identified in more than one tissue, as in the case of Foxj1, which is expressed in both the ciliated epidermis and the floor plate, and orchestrates motile cilia formation. Many of the genes correlating with foxj1 transcript abundance within the floor plate also correlated with foxj1 in the epidermis, and were highly enriched for ciliogenesis GO terms (Fig. 3.4F). Interestingly, we found that TFs deployed just once over the observed time series were more conserved in their expression pattern across the two-species compared to reused TFs (median correlation 0.72 vs. 0.48 in frog; 0.64 vs. 0.36 in fish). This may indicate that reused TFs are more commonly rearranged across tissues during evolution to generate novelties, whereas TFs whose field of expression is established just once might have more conserved regulatory functions.
Fig. 3.4: TF reuse is pervasive in vertebrate development. (A) Progressive programming of cell identity through sequential TF activation. (B) Programming of cell identity through combinatorial reuse of TFs. TF A is reused (red) in combination with TF B to generate a new fate. (C) Half of differentially expressed (DE) TFs are induced more than once in early frog and fish development. (D) Reused TFs increasingly dominate new TF expression during fate choices over time. (E,F) Reused TFs correlate with context-dependent (blue; off-diagonal) or conserved (red; on-diagonal) gene expression modules: (E) Pax8 correlates with different genes in the otic placode and the pronephric mesenchyme. (F) Foxj1 correlates with ciliogenesis genes in both the floor plate and in ciliated epidermal cells.
3.5 Discussion

In this chapter we looked for general features of development that might be difficult to appreciate from examination of individual lineages. In both Xenopus and zebrafish, we observed a transition from pervasive multilineage gene expression during gastrulation, to more specific and combinatorial deployment of differentiation programs later. In particular, newly induced transcription factors increasingly demarcated multiple independent cell fates in the embryo after gastrulation, as transcriptional states within each germ layer are further specialized. The drop in the frequency of multilineage priming, and the increase in combinatorial TF deployment, co-occur with the formation of the first discrete embryonic cell states. In Xenopus, this transition appears to be particularly switch-like, occurring between stages 12 and 14, just 2.5hrs apart in time. Curiously, the timing of appearance of distinct cells states correlates precisely with that of cell fate commitment in Xenopus (80,81). The global nature of this transition is reminiscent of the midblastula transition, that occurs in stage 8 Xenopus embryos, when the gradual titration of DNA-to-histone ratios during blastula stage cleavage cycles abruptly drives transcriptional activation, a longer cell cycle, onset of asynchronous cell divisions, and the start of cell motility across the embryo (208,209). We do not know whether any single mechanism coordinates the rapid appearance of multiple cell types in the post-gastrula embryo, however. A candidate process could be the formation of heterochromatin domains, which occurs on the appropriate timescale (4,111), and could facilitate both the refinement of lineage program conflicts (108 and the combinatorial re-use of transcription factors by restricting their targets in a tissue-specific manner (110,113).

3.6 Materials and methods

Cluster continuity analysis

We used the kurtosis of the cell distance distribution as a metric of cell clustering (Fig. 3.1). Specifically, the metric $c_i$ is calculated for each cell $i$, at each timepoint, as the $\ln(kurtosis)$ of the distribution of Euclidean dis-
tances from cell i to every other cell at the same timepoint, in the principal component latent space defined above in section “Visualization of data using PCA-tSNE and SPRING”. The metric serves as a continuity index because the appearance of discrete clusters leads to a bimodal or multimodal distance distribution for each cell, which increases the kurtosis. This analysis was performed on replicate 1 data.

**Transcription factor (TF) re-use analysis**

To identify TFs that were activated multiple times during early development, we used a graph based approach to search for TFs with multiple non-overlapping expression domains on the cell state tree. The cell state tree represents a directed graph. The general approach is (1) to binarize the gene expression for each gene on the cell state tree; and (2), to count the number of connected components of the graph after discarding all nodes that are non-expressing. This simple approach is elaborated with additional steps in order to (a) merge connected components that are close on the cell state tree, to conservatively allow for noise-induced “Gene-off” events; (b) filter out connected components consisting of just a single state or collections of isolated states due to merging, as these may be due to noise. We analyzed the list of differentially expressed (DE) TFs (the intersection of a TF list from and the DE gene list described above, filtered to exclude broadly expressed genes defined by a skewness over cluster averages less than 2). Then, for each gene, the detailed steps are as follows: (1) Gene expression is binarized on the cell state tree, classifying a cell state as ‘gene-on’ for a given gene if: i) the average expression was >x UMI/cell; ii) >y% of cells had detectable expression. (2) Additional edges are added to the cell state tree, connecting each cell state to all other cell states within z fate splits apart, i.e. states that can be accessed by walking on the graph and crossing no more than z-1 nodes that have >2 outgoing edges. (3) All “gene-off” states and their associated edges are discarded from the tree, to generate a graph consisting of “gene-on” cell states only. (4) A connected component of the “gene-on” graph was defined as robust if more than 50% of nodes in the component had at least one neighbor on the original cell state tree that is ‘gene-on’. Connected components that were not robust were discarded. (5) The number of connected components is the number of times a gene is activated on the cell state tree. The parameters x,y,z
determine the sensitivity at which a gene is considered expressed, and the resolution of domains of expression on the cell state graph. We chose values for these parameters that maximized the accuracy of calling a TF as single-use or re-used for a “ground truth” training set of 20 TFs that were manually inspected to be expressed only once, or more than once. The values that achieved maximum accuracy (18/20 TFs called correctly) were $x=0.1$ UMI/cell (500 transcripts per million (TPM)); $y=15\%$ cells; $z=3$ fate splits. The re-used TF analysis was performed on replicate 1 cells embedded in version 1 cell state tree. Results were tabulated in Additional Data Table S5 online.

Multi-lineage priming analysis

To systematically identify occurrences of multi-lineage priming (MLP), we searched for genes that were robustly co-expressed in the same cells at one point in time, and then still expressed but in disjoint sets of cells at a later point in time. Further, the latter cells are required to be descendants of the former cells in the cell state tree. To this end, we first classified pairs of differentially expressed genes in each of the fate splits on the tree as MLP+ or MLP-. Each individual gene was then deemed MLP+ with respect to a given fate split, if it contributed to at least one MLP+ pair at the fate split. Each gene pair was classified as MLP+ at a fate split as follows. For each sub-tree, we define a co-expression index for each time point, as the number of cells co-expressing both genes at the time point, divided by the number of cells expressing either gene. A gene pair was then classified as MLP+ for the fate split if: i) the maximum value of the co-expression index was $x$ times higher than the minimum value for the subtree, with the max happening earlier in time than the min; ii) the maximum value was higher than $y$; iii) the minimum value was lower than $z$. For Xenopus, we used $x=4$, $y=0.1$, $z=0.05$. For Zebrafish, we used the same parameters except $y=0.2$, to account for higher depth of sequencing. Note that the co-expression index is highly sensitive to the sequencing depth of the scRNA-Seq data, through the expression drop-out rate. Therefore, the absolute %MLP+ genes is an underestimate of the true value, but the qualitative trend in the %MLP+ genes over time is reliable, given a constant sequencing
depth. The main result reported in the main text – a global trend of decreasing MLP over time – was insensitive to 2-3 fold changes in each parameter. Results were tabulated in Additional Data Table S6 online.
Nature uses only the longest threads to weave her patterns, so that each small piece of her fabric reveals the organization of the entire tapestry.

Richard P. Feynman

4

Differentiation plasticity of stem cells
4.1 Chapter overview

In this chapter, we switch gears from the analysis of early embryos to focus on differentiation of stem cells in culture. There are two main questions: i) are cell states sensitive to the method used to generate them? And ii) can the same cell state be reached by multiple distinct differentiation paths? These questions are timely in the field of directed differentiation of stem cells and regenerative medicine. The advent of cell reprogramming has raised questions about whether artificial strategies to interconvert cell states, such as by transcription factor overexpression, is able to produce cell types that reliably reach the same mature terminal state. The questions are also fundamental in the sense that they address the underlying plasticity of differentiation programs. As we saw in chapter 2, orthologous cell states can in certain cases be reached by different differentiation trajectories across species. The experimental access of the stem cell system allows us to test how differentiation paths relate to terminal states in a controlled setting.

Our experimental strategy focuses on the conversion of mouse embryonic stem cells (mESCs) into motor neurons (MN), by two very different protocols. Motor neurons are a good mature cell state to focus on, as their development in the embryo and in culture is relatively well characterized. In comparing the two protocols, we focus on three main questions summarized in Fig. 4.1. In addressing them, we provide evidence at the single cell level that differentiation is path-independent, by directly observing two alternative differentiation trajectories from the pluripotent state into a functional motor neuron state.

*Parts of this work were first published as:* Briggs, J. A., Li, V. C., Lee, S., Woolf, C. J., Klein, A., & Kirschner, M. W. (2017). Mouse embryonic stem cells can differentiate via multiple paths to the same state. Elife, 6, e26945.

*Author contributions:* The results presented in this chapter result from an equal collaboration between myself and Victor Li. Victor was responsible for the stem cell experiments and generated Fig. 4.2A, Fig. 4.4C, and Fig 4.6. The scRNA-seq experiments and data analysis and the remaining figures were my own
Fig. 4.1: Summary of research questions. a) Direct programming could in principle skip progenitor states in one of two ways. It may either transition from early states directly into later states through a ‘short-circuit’ of the natural lineage, or it might utilize an alternative path involving new and potentially abnormal or unstable intermediate states. In the schematic, a conceptual depiction of the natural lineage is shown first, followed by two lineages that bypass the i2 intermediate from the normal lineage, either through short-circuit or through a new alternative intermediate. b) DP might diverge from the natural lineage only briefly, or it might access an entirely distinct path. In the schematic, three different possibilities are shown: i) DP bypasses specific early intermediates, and then converges to the final state through a conserved sequence of terminal cell state transitions; ii) DP transitions through conserved early stages of differentiation, before diverging into an alternate path that converges separately to the final state; iii) DP takes an entirely distinct path with no resemblance to the natural lineage. Dashed lines represent a potential direct transition while solid lines represent an alternate path (as shown in the left panel). c) It is possible that the price paid for taking an alternate differentiation path during direct programming is that cells retain subtle gene expression defects, thus converging only partially to the desired final state. Alternatively, convergence could be near perfect which would imply that differentiation into mature cell states is at least partially history independent. In the schematic, circles are an abstract representation of gene expression state and are used to represent the extent of overlap between MNs generated by either DP or by the natural lineage.

work. Victor and I wrote the initial manuscript together. Marc Kirschner and Allon Klein helped to design the study and edit the manuscript. S. Lee performed the electrophysiology experiments in C. J. Woolf’s lab. Esteban Mazzoni shared the transcription factor cassette containing ES cell line used in the reprogramming experiments that follow.

4.2 Dissection of two MN differentiation protocols by scRNA-seq

We compared two in vitro differentiation protocols that convert mESCs into spinal MNs, with the goal of identifying and comparing alternate differentiation trajectories between the same start state and similar – if not the same – end states. The first, standard protocol (SP) is a widely used method that attempts to drive cells into a spinal MN state through sequential addition of developmental signals that normally guide
MN differentiation in the embryo (Fgfs, Retinoic Acid, and Sonic hedgehog)(Fig. 4.2A) \textsuperscript{23,35}. Many variations of this method exist, including protocols that use 2D or 3D culture formats, serum or serum-free basal culture medias, and even partially different cocktails of growth factors. Yet, each variant is believed to shuttle cells through the same sequences of intermediate states \textsuperscript{23}, namely those that exist in the embryo. As a contrast to SP, we used an alternative direct programming (DP) strategy as our second method. The DP protocol drives cells toward a spinal MN fate by forcing the expression of spinal MN transcription factors (Ngn2+Is1+Lhx3)\textsuperscript{183,196} in cells grown in a simple growth factor free medium\textsuperscript{160}. Previous studies have shown that this DP protocol does not induce key marker genes of some embryonic MN intermediate states; for this reason DP was among the first differentiation protocols for which it was explicitly argued that a distinct differentiation process may be occurring\textsuperscript{183,196}. Our focus in this study was to understand how a non-embryonic differentiation trajectory differs from the SP, which drives cells through an embryo-like sequence of states (Fig. 4.1).

We used single-cell RNA sequencing (InDrops)\textsuperscript{143} to profile the differentiation in both protocols over time (Figs. 4.2B,C). Single-cell transcriptomics was our method of choice because of its ability to parse cell states and differentiation trajectories within populations that are not pure or contain rare intermediates, as expected for both DP and SP\textsuperscript{20,245,250,286,287,288}. We initially profiled 4,590 single-cell transcriptomes sampled from early (day 4/5) and late (day 11/12) timepoints for each protocol, and also used our previously published data from 975 mES cells as a day 0 reference\textsuperscript{143}. To visualize the combined single cell time course data for each protocol, we followed the same procedure described in Klein et al.\textsuperscript{143}: we identified highly-variable genes in each data set, retained non-noise dimensions after a principal component analysis on the standardized gene expression scores, and then employed t-distributed stochastic neighbor embedding (tSNE)\textsuperscript{202} to visualize the data (Fig. 4.2B-C; further details in Supplementary Methods). The most immediate feature revealed by tSNE was a gene expression continuum for each protocol, which correlated with chronology, suggestive of a differentiation trajectory (Fig. 4.2B-C, inset). Several disconnected cell clusters were also visible outside of the main trajectory.
By examining gene expression on the continua and disconnected cell clusters, we annotated the observed cell states produced in DP and SP. We applied unsupervised density gradient clustering (DBSCAN) to each tSNE map to partition cells into clusters, as described in Macosko et al\textsuperscript{73}. This approach robustly partitions each of the disconnected cell clusters, but can also identify boundaries between states within continua, although the latter are less reliably defined. The cells states identified in this manner possessed unique transcriptional signatures, including expression of specific transcription factors, signaling molecules, and structural genes such as synaptic components and neurotransmitters. To annotate each cell state cluster, we identified marker genes for each subpopulation, and used prior knowledge about the characteristics of these genes to interpret the identity of each state. The criteria used for each state are provided in Fig. 4.2-table supplements 1-2 online, and the expression of specific marker genes is shown in Supp. Figs. A14,15.

These cell state annotations reinforced our interpretation of the main gene expression continua in DP and SP as representing differentiation trajectories. In addition to correlating with chronology, the ordering of cell states appeared sensible in both cases. Each trajectory begins with mESCs (Pou5f1+/Esrrb+/Nanog+), passes through familiar neural progenitor states (including for example neural progenitors (NPs) expressing Sox1, Pax6, and Lin28), progresses to early motor neurons (EMNs) (expressing Mnxi, Isl1, Isl2, Nefl, Nefm, Tubb3, and Map2), and terminates at a mature, or late, MN state (LMN) (downregulating MN progenitor genes Mnxi, Isl1, and Isl2, while upregulating genes indicative of neuronal maturation Robo3, Onecut2, Syn1, Vacht, and Gabri) (Fig. 4.2B-E). Cells along both trajectories overlap between consecutive time points, which would be expected due to asynchrony in differentiation. This was reassuring to us that our experiments sampled all intermediate states. Below we define and compare the specific cell-state transitions that occur during MN differentiation in DP and SP, after first examining the efficiency of MN generation in each protocol.
Fig. 4.2: Dissection of DP and SP motor neuron differentiation strategies using InDrops single cell RNA sequencing. 

a) Summary of the direct programming (DP) and standard protocol (SP) differentiation strategies. 
b-c) tSNE visualization of single cell RNA sequencing data from each differentiation strategy. Timepoints are shown inset. Cell state clusters are color coded and annotated with their identities: ESC = embryonic stem cell; NP = neural progenitor; PNP = posterior neural progenitor; PVNP = posterior and ventral neural progenitor; MNP = motor neuron progenitor; EMN = early motor neuron; LMN = late motor neuron; CN = cortical neuron; DSCN = dorsal spinal cord neuron; G1 = glia type 1; G2 = glia type 2; Mus = muscle; Meso = mesoderm; Endo = endoderm; Oligo = oligodendrocyte; Astro = astrocyte; Stro = stromal. Arrows inside the bounded area indicate the hypothesized cell state progression during MN differentiation for each method. 
d-e) Expression heatmap of marker genes used to identify each subpopulation. Colors and annotations for the subpopulations are the same. 
f-g) Subpopulation abundances for each protocol over time. DP has significantly high MN production efficiency than SP (66% vs. 9%) at the late timepoint. MN = EMN or LMN; NP = NP, PNP, or PVNP; Other = everything else. Colors and labels match b-e).
4.3 Dynamics of MN Generation versus off-target Differentiation

Our single cell data allow us to quantify the dynamics of MN generation in DP and the SP over time, by counting the percentage of cells in each state at each time point. Previous reports have claimed up to 50% MN differentiation efficiency for the SP \cite{31,39}, and up to 98% efficiency for DP (after excluding cells that failed to activate the transgene)\cite{38}. However, both of these studies identified MNs as Mnx1+ cells, and for SP this was measured as GFP+ cells in an Mnx1:GFP reporter system. Since GFP is a stable molecule, and Mnx1 is first expressed in early committed MN progenitors, this definition includes cells from early MN progenitors (between PVNP and MNP in our classification) through to LMNs. If we apply these criteria to our day 5 SP data, we find a similar SP efficiency of 42.2%, and an efficiency of 66% for the DP protocol (not pre-filtering for transgene activating cells).

Using single cell data, it is possible to construct more refined criteria for cell states during MN differentiation, that resolve between the least mature MNPs, maturing EMNs, and mature LMNs, by focusing not on a single marker gene, but on the entire transcriptional state (Fig. 4.2B-E; Fig. 4.2-table supplement 1-2 online). As EMNs mature into LMNs they downregulate progenitor markers including Mnx1, and upregulate markers of terminal differentiation. For DP, MN production was observed as early as day 4 (17.8% EMN), and increased over time to 3.4% EMN and 62.7% LMN by day 11 (Fig. 4.2F). A minority of off-target neuron subtypes, glia, mesoderm, and endoderm cells were also identified, together accounting for <25% of the total population at day 11. We cannot rule out that these fates may emerge from cells that failed to activate the DP transgenes. In the SP, by contrast, we observed significantly lower efficiency in production of the most mature MN states. The SP population contained 18% EMNs and 1.1% LMNs at day 5, and just 6.5% EMNs and 2.6% LMNs at day 12 (9.1% total) (Fig. 4.2G). These figures are lower than those obtained from Mnx1-GFP scoring, because they do not include the earliest committed MN progenitors, which are yet to downregulate progenitor genes such as Mnx1 and Isl1 or to upregulate neuronal genes such as Tubb3 and Map2, and because they distinguish between early and late MNs by their expression of terminal differentiation markers.

83
such as Syn1, Mapt, Celf4 or Gabri. Our single cell classifications therefore capture varying degrees of maturity among differentiating MNs.

An unexpected result from this analysis was that, in the SP protocol, off-target populations increase in fractional abundance to overshadow the intended MN products. Off-target products included oligodendrocytes (7.2%), astrocytes (8.6%), muscle (30.6%), and stroma (8.9%), that together accounted for 53.3% of the population by day 12. The increase in these populations possibly reflects the ability of cell states such as oligodendrocytes and astrocytes to proliferate, which could progressively dilute the initial output of post-mitotic MNs, severely reducing MN generation efficiency. To test this idea and to gain additional EMN and LMN transcriptional data, we profiled an additional 1,372 single cell transcriptomes from a replicate SP experiment at day 10 (Fig. 4.3; Supplementary methods). At this time point, the SP population contained 34.4% EMNs and 14.9% LMNs, which was higher than at either day 5 or day 12. This is consistent with the hypothesis that SP has an initial output of MNs, which may then be diluted by proliferating off-target lineages.

4.4 The DP differentiation trajectory lacks Nkx6-1+/Olig2+ intermediate states

What are the differentiation paths taken by each protocol? In the SP differentiation path, cells transit through seven states (Fig. 4.2C,E). These state transitions parallel patterning events in the embryo: cells first commit to the neural lineage (NP; Sox1+/Sox2+), then are posteriorized (PNP; Sox3+/Hoxb8+/Hoxd4+), ventralized (PVNP; Nkx6-1+/Olig2+), enter the committed MN progenitor state (MNP; Mnx1+), and then mature through early (EMN; Mnx1+/Tubb3+/Map2+) and late (LMN; Mnx1-/Tubb3+/Map2+/Syn1+) MN states. This is not a surprise as the growth factor cocktail defining this method was designed to reflect the signaling events taking place in the embryo. By contrast, we found that the path produced by DP was condensed relative to the SP path (Fig. 4.2B,D), consisting of only four apparent states as opposed to seven. After neural commitment (NP; Sox1+), cells immediately began expressing committed MN markers (EMN; Mnx1+/Tubb3+). This immediate transition was also registered by time course qPCR measurements for a
panel of MN genes across bulk populations (Supp. Fig. A.16).

A major difference between the DP and SP cells states is the seemingly complete absence of typical spinal embryonic intermediates (PNP or PVNP) in DP. These states are normally recognized by expression of the marker genes Olig2, Nkx6-1, Hoxb8, and Hoxd4, but these genes were not detected by scRNA-Seq in DP intermediates. Olig2 is necessary for MN development in embryos\textsuperscript{278}, indicating that DP drives differentiation from ESC into MNs through a fundamentally new route.

To more confidently demonstrate that DP does not transit through an Olig2+ state, even transiently, we conducted a high-sensitivity gene expression analysis. In SP, our scRNA-Seq data reveals that Olig2 expression peaks (in the PVNP cells) at a level 5.4-fold higher than Gapdh, a constitutively expressed ‘housekeeping’ gene. We performed a dense qPCR time course measuring the expression of Olig2 and a panel of MN genes every day during DP across populations of \texttimes{}106 cells. These measurements showed that committed MN markers were upregulated immediately following early neural progenitor genes in DP, but Olig2 was not detected at any time point during DP (no qPCR signal at CT>40), showing that the average expression levels are \texttimes{}106-fold or more lower than Gapdh (Supp. Fig. A.16). Accordingly, on any day, the fraction of cells expressing Olig2 at levels seen in SP was less than one in \texttimes{}106. Given that the total protocol duration was 11 days, this would require that an Olig2 transitional state would last less than 1 second (11 days/\texttimes{}106), far shorter than the lifetime of mRNA molecules in a cell. Therefore, there is no significant Olig2+ transitional state in DP. This conclusion holds even after allowing for the possibility that Olig2 might act through just a single mRNA transcript. From the qPCR data, the fraction of cells expressing Olig2 at 1 molecule per cell cannot be more than 0.1\% (see Supplementary Methods), and the corresponding lifetime of Olig2+ cells would correspondingly be 15 minutes (11 days/\texttimes{}103), still inconsistent with the lifetime of a single mRNA molecule. These results are consistent with a previous study that reported Olig2 absence during DP\textsuperscript{181}.
4.5 Bifurcation and convergence of the DP and SP trajectories

Since the DP omits spinal embryonic intermediates characteristic of the SP path, there must be one of two possible trajectories. Either DP must discontinuously transition from an early neural progenitor into a MN, or it must transit through alternate intermediate state(s). To determine which of these possibilities was the case, we employed a data visualization technique called SPRING to directly compare the topology of both paths. While tSNE is a powerful method for identifying discrete cell states, SPRING provides a complementary description emphasizing continuum gene expression topologies. SPRING builds a k-nearest-neighbor graph over cells in high-dimensional gene expression space, and then renders an interactive 2D visualization of the cell graph using a force directed layout. This representation revealed that the DP and SP trajectories overlap during early neural commitment, but that they then bifurcate and take distinct paths that converge independently to the same MN state (Fig. 4.3A). The dynamics of gene expression over these trajectories resembled the behavior inferred using tSNE, with DP omitting intermediate progenitor genes following its bifurcation from the SP path (Fig. 4.3B).

The bifurcation and subsequent convergence of the two differentiation paths is also revealed by two other complementary analyses. Pairwise cosine similarities between the cell states from both trajectories (Fig. 4.3C; Supplementary methods) indicate similarities between the early states (ESC and NP; cosine similarity > 0.69) and late states (LMN; cosine similarity ≈ 0.47), but not the intermediate states (PNP, PVNP, and MNP; cosine similarity < -0.15, -0.03, and 0.04 respectively). We also assigned every individual cell along the DP path to its most similar cluster in the SP path using a maximum likelihood method (Fig. 4.3D; Supplementary methods). This showed that it was virtually impossible to find a single cell resembling the SP intermediate progenitors in the DP approach. Similarity was again observed only at the early and late states.
**Fig. 4.3: DP and SP induce distinct differentiation paths to the motor neuron state.**

a) Visualization of differentiation paths for both protocols using SPRING reveals two paths to the same state. DP and SP overlap during early neural commitment, but then bifurcate and converge to the same final MN state separately. Reds = direct programming path; Blues = standard protocol path; arrows indicate hypothesized differentiation trajectories. Cell states are colored and labeled by according to their definition in Fig. 4.2 for comparison.

b) Gene expression of key marker genes along each differentiation path confirms exit from the pluripotent state (Oct4) and progression towards the MN state (Tubb3, Stmn2) for both protocols. Only the SP upregulates Olig2 and Nkx6-1, which mark important MN lineage intermediates in the embryo; this occurs following the bifurcation of both paths. Expression in cells from each sample are colored using either a blue (ESC), red (DP), or green (SP) colormap to allow tracking of each path separately.

c) Pairwise cosine similarity of cell state centroids. Note early and late similarity of states, but prominent differences during intermediate state transitions.

d) Every individual cell of the DP trajectory was assigned to its most similar SP state using a maximum likelihood method. DP cells map to early and late but not intermediate states of the SP cell state progression.
4.6 DP transitions through an novel intermediate state

The bifurcation of the SP and DP trajectories leads to different intermediate cell states. A total of 26 transcription factors (TFs) are differentially expressed between the DP and SP intermediate states (Fig. 4.4A). A majority of these (61%) were involved in an anterior-posterior positional gene expression axis. The SP intermediates were enriched more than 6-fold for nine posterior and spinal TFs including Olig2, Nkx6-1, Lhx3, and six posterior Hox genes with a corrected p-value < 0.001. Each of these TFs is expressed in embryonic MNs. By contrast, the DP intermediates were enriched for seven forebrain TFs including Otx2, Otx1, Crx, Six1, Dmrt2, Zic1, and Zic3 at the same stringency, despite the absence of MNs in the forebrain of embryos. Anterior gene expression was previously observed through bulk measurements of DP\textsuperscript{18}, and our results reveal that it occurs within a specific subpopulation of cells in the process of differentiating into MNs.

The positional gene expression signature that characterizes the DP intermediate state appears to be transient. Forebrain gene expression is upregulated along the DP differentiation path, as cells exit the early NP state into the EMN intermediate state (Fig. 4.4B). This transition is accompanied by the downregulation of proliferation-associated genes (Fig. 4.4B; Supp. Fig. A.18). By the time cells exit the EMN state and transit into the more differentiated LMN state, they downregulate forebrain genes and replace this abnormal positional signature with a spinal Hox expression signature characteristic of normal MNs (Fig. 4.4B). Thus cells converge to the MN state in positional as well as neuronal identity gene expression in the final stages of DP.

We validated the existence of the abnormal transitional state in DP in two ways using an Mnx1:GFP reporter cell line. Mnx1 serves as a useful reporter in this context since its expression is localized precisely to the distinct intermediate populations of each differentiation path (Fig. 4.3B). First, we performed bulk microarray comparisons of Mnx1+ DP and SP intermediates to compare their gene expression states by an independent method. The comparison confirmed the enrichment of forebrain TFs, and depletion of spinal progenitor and positional genes in the DP intermediates with just one exception – Zic1 was enriched in DP by our single cell comparison but in SP by microarray (Fig. 4.4A). Second, we isolated Mnx1+ DP intermediate cells
Fig. 4.4: During DP cells transition through an abnormal intermediate state with a forebrain gene expression defect. a) The DP (red, EMN from Fig. 4.2) and SP (black, PVNP+MNP+EMN from Fig. 4.2) differentiation paths diverge into distinct intermediate states following early neural commitment. These distinct populations were compared by in silico differential expression analysis using single cell data (bottom left), or by FACS purification of the intermediate populations using and Mnx1::GFP reporter cell line followed by microarray analysis (bottom right). Fig. 4.3B shows that Mnx1 expression localizes to the distinct intermediate populations of both protocols, making it an appropriate target for their purification. Differential expression analysis revealed a total of 26 differentially expressed TFs between the intermediate populations (>6-fold differentially expressed, p<0.001), of which (61%) were involved in either forebrain positional identity (Otx2, Crx, Zic1, Six1, Dmrt1, Otx1, and Zic3) and enriched in DP, or in posterior spinal identity (Hoxb8, Hoxb3, Hoxb2, Hoxb4, Hoxb5, Hoxd4, Lhx3, Olig2, and Nkx6-1) and enriched in SP. The independent microarray comparison confirmed this differences with only one exception - Zic1. It also revealed that the SP intermediates are enriched for additional spinal neural progenitor genes including Ascl1, Nkx6-2, Olig1, and Nkx2-2. b) Before forebrain genes are expressed, DP cells have a high proliferation score and low cell-cycle exit score (computed as the sum of a panel of cell-cycle-associated or tumor suppressor genes respectively). They reduce cell cycle gene expression as they enter the abnormal transitional state, and upregulate forebrain genes including Otx2 and Crx. This abnormal forebrain expression is shut off as cells exit the transitional state into the final MN state, characterized by Map2 and Vacht expression. The final transition into a MN state is also accompanied by upregulation of posterior positional genes including Hoxb4 and Hoxd3, thus correcting the transient forebrain positional expression defect. Expression in cells from each sample are colored using either a blue (ESC), red (DP), or green (SP) colormap to allow tracking of each path separately. c) Mnx1+ cells were isolated at day 3 of differentiation and cultured to determine their fate potential. By 12 days after replating most cells expressed motor neuron markers Map2 and Vacht. Scale bar = 100um.
on day 3 of differentiation, the earliest time point at which Mnxi+ cells appear, and prior to maturation of any Mnxi-GFP+ cells. We cultured this population for an additional 10 days, and validated by immunostaining for Vacht and Map2 that it indeed generates motor neurons (Fig. 4.4C). This experiment confirms that the abnormal transitional state in DP gives rise to MNs. The abnormal DP intermediate thus simultaneously has a forebrain positional expression state and possesses MN fate potential.

4.7 DP and SP end states closely match MNs from embryos

Given that the two protocols induce distinct – and in the case of DP, unnatural – differentiation paths, we were curious how their final products compared with primary MNs (pMN). We harvested MNs from the embryo of a Mnxi:GFP reporter mouse and performed inDrop measurements on 874 Mnxi+ cells that were FACS purified from whole E13.5 spinal cords. Though the majority of Mnxi+ sorted cells were MNs (73.8%, n = 645), this population also contained glia (20.1%), fibroblast-like cells (1.8%), and immune-type cells (1.2%; Fig. 4.5A; Supp. Fig. A.19). Using only the cells identified as bona-fide MNs, we compared the differentiating DP and SP cells to pMN by both global transcriptome similarity of cell states centroids, and a nearest neighbor analysis of single cells. Global transcriptome comparisons confirmed that each state along the DP and SP differentiation paths becomes progressively more similar to pMN (Fig. 4.5B). The clusters most similar to pMN were the LMN state from the DP protocol (cosine similarity = 0.60), and the LMN state from the SP (cosine similarity = 0.47). Since subsets of LMNs from DP and the SP might vary in similarity to pMN, we analyzed the similarity of single cells from all three experiments using SPRING, by embedding all three data sets onto a single kNN graph. We performed this analysis including all cells (Fig. 4.5C-i), and then including only EMNs, LMNs, and pMN (Fig. 4.5C-ii). Both approaches showed that pMN closely associate with the LMNs of both DP and SP. It was also apparent that DP and SP LMNs are themselves heterogeneous, with particular subsets associating more closely with pMN. Overall, a higher fraction of DP LMNs resembled primary MNs, as seen by calculating the fraction of cells in each state that had at least
one pMN nearest neighbor out of its 50 most similar cells (Fig. 4.5C-iii; 64% for DP, 6% for SP). DP LMNs therefore appear if anything more related to pMNs in gene expression than SP LMNs, despite their unusual developmental path.

The differences in similarity of DP and SP derived MNs to primary MNs can be understood in terms of differences in maturity and identity of the cells. As expected, both DP and SP EMNs were less mature than pMNs, as seen by expression of the MN progenitor marker Mnx1 (Fig. 4.5C-iv). Among the SP LMNs we detected multiple subpopulations. One subset appeared more anterior than the spinal E13.5 pMNs, as indicated by expression of the hindbrain marker Gata3; the second appeared less mature than pMNs as indicated by residual expression of progenitor genes such as Mnx1 or Nkx2-2; and the final population was closely associated with pMNs (Fig. 4.5C-iv). Most DP LMNs by contrast were closely related to pMNs. They lacked the residual progenitor gene expression seen in SP, and they shared with pMNs the expression of genes important to terminal MN differentiation (e.g. Ebf2 and Ebf3). Beyond individual marker genes, we also systematically compared DP LMNs, SP LMNs, and pMNs by pairwise differential gene expression analysis (Fig. 4.5D). Overall, both DP and SP LMNs show limited gene expression differences to pMNs. In both cases there were 11 differentially expressed TFs. Of these, half were positional Hox genes that mark the most posterior MN pool and were enriched in pMNs, indicating a more anterior identity of the DP and SP LMNs. The remaining TFs included Gata3, which marks a hindbrain subpopulation of SP LMNs as noted above, and other genes related to neuronal maturation including Neurod6, Ebf2, and Mid1. These comparisons confirm the broad similarity of DP LMNs and SP LMNs, to pMNs, but reveal subtle differences between these terminal states that relate to positional identity and state of maturation.

4.8 DP MNs have physiological morphology and function

Having established that MNs derived via both gene expression trajectories reach roughly the same MN transcriptional state, we wished to validate that their function and structural organization was also independent
Fig. 4.5: Both DP and SP differentiation trajectories approach the transcriptional state of primary MNs (pMNs), but DP does so with higher precision. a) tSNE visualization of 874 single cell transcriptomes from FACS purified Mnx1+ MNs from embryos reveals heterogeneity within this population. To make comparisons between DP and SP with pMNs we used only the subset of Mnx1:GFP+ primary cells in a bona-fide MN state. See Supp. Fig. A.19 for marker gene expression in each population. b) Comparison of average gene expression profiles for cell states along the DP and SP trajectories with pMNs. In both methods similarity increases as differentiation proceeds. Late DP states are the most similar to embryonic MNs. c) Projection of the reference E13.5 pMNs into the visualization from Fig. 4.3 revealed that pMNs closely associate with the terminal states of both DP and SP (i). Close examination of the terminal populations (EMN, LMN) from DP and SP compared to pMNs reveals heterogeneity representing state subtypes (ii). At a single cell level DP LMNs were the most closely associated with E13.5 pMNs; 64% of DP LMNs had at least 1 pMN nearest neighbor out of its most similar 50 cells compared to 6% for SP LMNs (iii). The subtypes present within terminal DP and SP populations could be annotated using marker genes. DP and SP EMNs express progenitor genes including Mnx1, along with Nkx2-2 and Nkx6-1 in SP only. The major SP LMN outgroup expressed Gata3, indicating a hindbrain identity. Both DP LMNs and pMNs shared expression of the terminal MN differentiation gene Ebf2. d) Systematic pairwise differential gene expression analysis between terminal DP and SP states and pMNs. Each panel is a volcano plot of differentially expressed transcription factors. Both DP and SP LMNs show limited gene expression differences to pMNs. The dominant differences are positional, with DP and SP LMNs lacking the most posterior Hox genes. Other expression differences are explained by differences in terminal state subtypes as shown in c). Differentially expressed genes were filtered for TFs with a corrected p-value < 0.001, expression ratio > 4, and minimum expression of 1umi/cell average.
of their distinct developmental histories. The SP has been characterized extensively as giving rise to functional MNs, so here we examined structural and functional characteristics following DP. We confirmed that selected protein content matches the mRNA markers by immunostaining for Tubb3, Map2, VACht, Isl1, and Hb9 (Fig. 4.6A). Tubb3 and Map2 were present, and VACht was seen at discrete puncta on the axons (suggesting localization to acetylcholine secretory vesicles). TFs Isl1 and Hb9 were localized in the nucleus. Finally, the GFP from the Mntxi-GFP reporter was activated and expressed in the cytoplasm. To test the functional properties of the DP MNs, we performed whole-cell patch clamp recordings. Depolarization induced single or multiple action potentials in current-clamp experiments (Fig. 4.6B). Depolarizing voltage steps induced fast inward currents and slow outward currents characteristic of sodium and potassium channels, respectively (Fig. 4.6C). Exposure to 500 nM Tetrodotoxin (TTX) blocked the inward current, indicating sodium channel involvement. We then tested whether our DP neurons would respond to neurotransmitters that act on MN (Fig. 4.6D). Exposure of the neurons to AMPA, kainate, GABA, and glycine (100 μM each) induced in each case inward currents similar to that seen in primary embryonic MNs. To see if the DP neurons could also form neuromuscular junctions, we co-cultured the neurons with differentiated C2C12 skeletal muscle myotubes and incubated them for 7 days. We observed clustering of acetylcholine receptors on the C2C12 myotubes near contact points with the DP neurons, which can be seen with alpha-bungarotoxin (α-BGT), which binds to acetylcholine receptors (Fig. 4.6E). We then observed regular contractions of some C2C12 myotubes that began after several days in co-culture (Fig. 4.6E, also see Fig. 6-video supplement 1 online). These contractions could be stopped by the addition of 300 μM Tébocu-curarine (curare), an antagonist of acetylcholine receptors, indicating that the contractions were induced by the acetylcholine release from the MNs. Similarly, we noticed that the DP MNs could induce contractions in DP muscle myotubes that we previously generated with MyoD (see Fig. 6-video supplement 2 online). These results confirm that DP MNs have the expected functional properties of bona-fide MNs.
Fig. 4.6: Validation that DP cells become functional MNs despite their abnormal differentiation trajectory. a) Immunostaining of MN markers in DP MNs confirming expression and correct subcellular localization of Tubb3, Map2, VACht, Isl1, and Hb9. DP MNs also: b) can fire single or multiple action potentials upon stimulation, c) show sodium and potassium channel currents, and d) are responsive to multiple MN neurotransmitters - AMPA, kainate, GABA, and glycine. e) Co-culture experiments show that DP MNs can induce clustering of acetylcholine receptors on muscle myotubes (indicated by α-BGT staining) and induce their contractions. The induced contractions are dependent on MN activity as they can be blocked by addition of the acetylcholine antagonist curare.
4.9 Discussion

We provide evidence at the single cell level that differentiation can proceed by multiple routes and yet converge onto similar transcriptional states. While cells differentiated by growth factor stimulation in the SP retrace the embryonic lineage, cells differentiated by terminal state transcription factors in DP take a dramatically different path. The DP path bypasses multiple intermediate progenitor-states that are produced in the embryo, and yet they still converge to the same discrete and recognizable MN phenotype. This convergence occurs via an abnormal intermediate state, and does not appear to involve a shared set of terminal cell state transitions. Moreover, as cells converge they manage to not only establish gene expression related to MN functions, but they also correct positional gene expression defects (exchanging forebrain for spinal gene expression) in the absence of external signals. We conclude that DP of mESCs into MNs occurs via a late bypass that involves alternative intermediate states not seen in the embryo, and that this new route converges near perfectly to the same final state (Fig. 4.1). Convergence into a MN therefore does not depend rigidly on the precise history of intermediate states through which cells differentiate.

This ’history independence’ of the final state is consistent with a dynamical view of gene regulation in which cell states correspond to ’attractor basins’, i.e. stable states of gene expression that are robust to modest perturbations. If attractor basins do not exist, the precision of the observed overlap between DP and SP MNs would require a special coincidence. The concept of cell states behaving as attractors has been proposed previously to explain several properties of blood cell types. There are at least two important corollaries of this behavior applied to motor neuron development. From a practical perspective, it is a common concern that DP methods may generate cell types with subtle defects due to their unusual developmental histories. Attractors would be robust to this vulnerability and indeed our results show that it is not necessary to recreate the precise sequences of steps taken in embryos to generate MNs with the highest similarity to primary cells. It could also hint at a mechanism that might help animal body plans evolve flexibly. Specifically, by decoupling the identity of mature cell state attractors from their developmental histories evolution would
be able to act on each independently. In principle this could contribute to evolvability by allowing mature cell states to be transposed onto new lineages in new body locations, without massively rewiring the internal differentiation circuit.

The mechanisms that define the MN attractor basin and allow the artificial DP trajectory to converge onto the correct final state are unclear. The MN state is thought to be stabilized by a network of self-reinforcing TFs, involving Mnr2, Mnx1, Lim3, Isl1, Isl2, and Lhx3, Ngn2, Myt1l, Nefl, and Nefm. DP aims to kick-start this network by activating a subset of important components. Yet, far from immediately activating this network, our data show that DP initially drives cells to differentiate into an early NP state through the same pathway as the SP trajectory, seemingly oblivious to the DP TFs, and then even activates non-MN genes in the transitional state. Understanding why the activation of the MN program lags behind TF induction may provide important clues into how the DP factors act. One possible source for a lag is that activating a complete neuronal program requires first activating additional core TFs (so-called ‘feed-forward’ circuitry). Indeed, recent studies have shown that Ebf and Onecut are activated by Ngn2 (one of the DP TFs), and that both are required to subsequently direct binding of Isl1 and Lhx3 (the other two DP TFs) to MN target genes across the genome during DP. A second possible source of lag is that extracellular signaling provides inputs that immediately affect cell state, but take time to sensitize cells to the DP factors. For example, signaling changes might activate DP TFs through post-translational modifications, by activating co-factors, or by inducing chromatin state changes. We have indeed observed that MNs are not generated if DP TFs are induced in cells cultured in pluripotency media, indicating a requirement for changes in signaling (not shown). Conversely, when mES cells are transferred to minimal media without inducing DP factors, they acquire a forebrain neural progenitor identity by default. This suggests that the early dynamics and abnormal forebrain/MN expression of the DP transitional state might in fact be driven by the signaling environment and not the DP TFs. These alternatives suggest future experiments to better resolve the mechanisms driving the DP, by re-mapping the trajectory induced during DP after changing signaling conditions, or the choice of DP TFs.
As a methodology, DP is significantly more efficient than the SP without loss of quality in the MN populations produced (Fig. 4.2F-G; Fig. 4.5; Fig. 4.6). The high-efficiency of DP most likely derives from both its more uniform experimental conditions as well as its more direct differentiation path. Experimentally, DP: relies on 2D rather than 3D tissue culture (as in the SP), minimizing uncontrolled cell-cell communication; forces every individual cell to express MN TFs from a genetically integrated construct, increasing uniformity; and employs a defined-media without growth factors that may minimize proliferation of competing progenitor states. The more direct differentiation path induced by DP might also itself increase MN conversion efficiency by minimizing error propagation through off-target fate choices. Long sequences of intermediate cell state transitions offer many sequential opportunities for off-target fate choices, which multiplicatively reduce efficiency, as compared to the more compact DP trajectory. Targeting terminal attractor basins through the shortest possible differentiation paths may prove to be a generally effective strategy for generating desired cell states, more quickly and more enriched in the desired product.

4.10 Materials and methods

ESC culture

The mouse ES cell line containing doxycycline-inducible Ngn2+Ish1+Lhx3 (NIL) and the Hb9-GFP reporter was graciously provided by Esteban Mazzoni. ESCs were cultured in standard media (DMEM with LIF + 15% fetal bovine serum) on 0.2% gelatin-coated dishes.

Differentiation into motor neurons: direct and standard programming

Twenty-four hours before starting differentiation, ESCs were trypsinized and seeded onto plates pre-coated with a mix of poly-d-lysine (100 μg/ml) and laminin (50 μg/ml) instead of gelatin for adherence. ESCs were counted by a Beckman Coulter Counter and seeded at a density of approximately 200,000 cells per well of a 6-well dish. At day 0 (ESCs), the media was switched from standard ES media to N2B27 media (Invitrogen).
Doxycycline was also added at 3 μg/ml starting to induce expression of the NIL transcription factors. Media was changed daily. For the standard programming protocol, the steps described in Wu et al\textsuperscript{35} were followed strictly. The Wu et al. protocol involves a period of neural induction using embryoid body culture before subsequent replating on laminin. The culture media is serum containing base media supplemented with specific cocktails of growth factors that change during the early, middle, and late stages of differentiation (Fig. 4.2A).

**Mouse embryonic motor neuron cultures**

The B6.Cg-Tg(Hlxb9-GFP)1Tmj/J mice (JAX\# 005029) were bred with C57BL/6J (JAX\# 000664) for embryonic motor neurons dissection. All animal protocols were approved the Institutional Animal Care and Use Committee at Boston Children’s Hospital. On gestational day 13 (E13), the female mice were anesthetized and all embryos were collected by caesarian section. Only GFP embryos were used for further dissection. The spinal cords were isolated and their meninges were removed. Each isolated spinal cord was dissociated by trypsin and mechanical trituration. After filtering the cells with 100 μm strainers, the cells were spun down and re-suspended in PBS, and subjected to flow cytometry. Cells were run through a 100 μm nozzle at low pressure (20 p.s.i.) on a BD FACSaria II machine (Becton Dickinson, USA). A neural density filter (2.0 setting) was used to allow visualization of large cells.

**Single cell transcriptomics using InDrops**

We dissociated differentiating mESC cultures using a 0.25% Trypsin 2mM EDTA solution (Gibco). Primary HB9+ sorted motor neurons were dissociated as above. Dissociated cell suspensions were verified to be monodisperse and of viability >95% using a coulter counter with trypan blue staining (BioRad). We then performed droplet-based barcoding reverse transcription (RT) reactions and prepared massively multiplexed sequencing libraries using InDrops as described in Klein et al\textsuperscript{43}. Briefly: cells, lysis and RT reagents, and barcoding primers attached to hydrogel beads are combined in nanoliter sized droplets suspended in an oil
emulsion using the InDrops microfluidic device. A barcoding RT reaction is then carried out in the droplet emulsion, uniquely labeling the RNA contents of every cell using a cell barcode and unique molecular identifier (UMI). Following RT, barcoded emulsions are split into batches, and the emulsion is broken. Combined material is amplified into a nanomolar Illumina sequencing library through a series of bulk reactions: second strand synthesis, in vitro transcription (IVT), fragmentation, RT2, and a final low cycle number PCR. The majority of the amplification takes place during IVT ensuring uniform library coverage. Single cell libraries were then sequenced on either the Illumina HiSeq or NextSeq platforms. Reads were demultiplexed using an updated version of the custom bioinformatics pipeline described in Klein et al. A python implementation of this pipeline is now publically available on GitHub. Briefly, it filters for reads with the expected barcode structure, splits reads according to their cell barcode, aligns them to a reference transcriptome (we used GRCm38 with some added mitochondrial genome transcripts), and then counts the number of different UMIs appearing for each gene in each cell. The final output is a counts matrix of cells vs. genes that we loaded into MATLAB for further analysis.

**Single-cell data clean up: minimum expression, total count normalization, stressed cell removal**

Before performing the analyses described below, three steps were taken to ensure that the data was of high quality. First, we required all cells to have at least 1000 UMIs detected. This removed any signal potentially coming from empty droplets. Second, data were total count normalized to ensure differences between cells were not due to technical variation in mRNA capture efficiency or cell size. Finally, cells that had a high stress gene signature were excluded from analysis. Stressed cells were initially recognized as a small percentage of cells (<10%) that clustered apart from everything else and specifically expressed very high levels of a mitochondrial gene set that is associated with cellular stress. Masks that convert raw counts into our filtered set are provided online.

**Visualization of single-cell data using tSNE**
To visualize high-dimensional single cell data, dimensionality reduction is essential. We chose to implement tSNE as described in Klein et al\textsuperscript{143}. The core steps are summarized as follows. Steps 1-2 precede tSNE, and focus the algorithm on genes that best describe differences between cell populations.

1. Extract the top 1000 highly variable genes. We do this using a statistical test derived specifically for InDrops data (Klein et al.)\textsuperscript{143}.

2. Extract principal variable genes. Principal variable genes are a subset of the highly variable genes from step 2 that we find best describes the cell population structure. The steps to find principal variable genes are:
   
   (a) Perform PCA using the top 1000 biologically variable genes.
   
   (b) Identify the number of non-trivial principal components. This is done by comparing the eigenvalue of each principal component from a to the eigenvalue distribution for the same data after being randomized. Only principal components with eigenvalues higher than those observed on random data are retained.

   (c) Extract genes that contribute most highly to these principal components by imposing a threshold on the gene loadings for each non-trivial PC.

3. Perform tSNE. We used the MATLAB implementation of tSNE from Van de Marteen et al\textsuperscript{592}. As input, we supplied z-score normalized principal variable genes, and asked tSNE to perform an initial PCA to a number of dimensions equal to the number of non-trivial principal components in step 4. tSNE then takes cells embedded in this PCA space and nonlinearly projects them onto two dimensions for visualization.

---

Subpopulation analysis

The first goal of our single cell analysis was to describe the identity and proportions of cell states generated by each of two motor neuron differentiation strategies. For this purpose we found that good results were
achieved using local-density gradient clustering on the 2D tSNE representation of the data. This approach provided a clear and natural cell state classification that was well aligned with prior knowledge about marker gene expression domains. The steps we took are summarized as follows:

1. Perform tSNE (as above) on cells pooled from all timepoints for each protocol (e.g. Fig. 4.2B).

2. Apply local density gradient clustering to define cell states.

3. Identify genes specifically expressed by each cell state, and use prior knowledge on their expression domains to generate a cell type annotation (e.g. Fig. 4.2D).

4. Quantify the fraction of cells in each state at each timepoint (e.g. Fig. 4.2F).

We identified genes that were specifically expressed by each subpopulation through pairwise t-tests and visually inspected their expression over the tSNE embedding. In Fig. 4.2D-E of the main text, we show the z-score normalized expression of a selection of marker genes that we used as a heatmap. Z-score normalization preserves differences between populations while putting the expression level of every gene on the same scale. In Supp. Figs. A14-15, we show the un-normalized expression of each of these genes individually so that the reader may compare.

**Initial identification of differentiation trajectories**

During our subpopulation analysis we observed, for both protocols, a continuous progression of cells that spanned the initial ES cell state through the early and late differentiation timepoints. This progression was punctuated by familiar progenitor states that were ordered in a way that was consistent with known events in motor neuron differentiation (Fig. 4.2D-E), and was correlated with chronology (Fig. 4.2B-C inset). We interpreted these progressions as differentiation trajectories. Each is reconstructed from three population snapshots (day 0, day 4/5, and day 11/12). Because differentiation in vitro is asynchronous, the single cell data overlapped from one timepoint to the next. We deduce from this that intermediate states were not missed
due to the spacing of our timepoints. We also validated that important intermediate genes were not detected over a densely sampled timecourse using qPCR (see below).

**Differential gene expression analysis between cell states from single-cell data**

We identified differentially expressed genes between cell states by using two-tailed t-tests with a multiple hypothesis testing correction. We defined differentially expressed genes conservatively at a FDR of 5% and a significance level of \(p<0.0001\). We only considered genes where at least one of the states being compared had \(\geq 10\) cells with non-zero expression. To find marker genes of a population we asked for genes that were enriched in that population versus everything else. For several comparisons we restricted our analysis to Riken transcription factors. This list contains 1,500 genes with manually annotated transcription factor activity. We represented differential expression data using volcano plots, and colored the expression intensity of each gene using a colorbar; the mean of the higher expressing state was used.

**Integration of replicate SP EMN/LMN cells into original SP trajectory**

Due to the low efficiency of EMN / LMN production in our initial SP experiment, we performed a replicate experiment with the goal of extracting additional SP terminal state cells. We repeated SP differentiation and performed InDrops on an additional 1,372 cells sample at day 10 of differentiation. To identify EMN or LMN cells within this population we: projected replicate SP cells into the PC-space from the original experiment; calculated a Euclidean distance matrix between all cells; extracted replicate SP cells which had \(\geq 3\) of their 5 nearest neighbors belonging to EMN or LMN clusters; then finally assigned cluster identities to these cells such that cells were called EMN if a majority of their neighbors were EMN, or otherwise were called LMN (majority LMN nearest neighbors or equal EMN and LMN nearest neighbors). This final rule slightly favors LMN over EMN cluster assignment but this helps to counter the low LMN efficiency of the original experiment, which reduces the possible nearest neighbors available during label assignment. For all comparisons of DP and SP we pooled EMN and LMN cells across the two replicate experiments for SP, improving the statistics of the end state comparisons.
Comparison of DP and SP differentiation paths

Comparison method I: Visualization of alternate differentiation trajectories using SPRING. After our initial identification of the differentiation paths for the standard protocol and for direct programming we wished compare their routes. One of the most powerful ways to begin addressing such a problem is to simply visualize the data. Yet, we found tSNE gave unclear results for a direct comparison of the paths; visualizing both protocols together resulted in some mixed clusters, and some distinct clusters, but no overall coherence to the representation as we had found looking at one or the other trajectory separately. The limitations of tSNE for analyzing continuous processes are well known.

We therefore turned to a new method developed in parallel to this work in our lab called SPRING, that in our experience does better in analyzing continuous processes in single cell data. SPRING has four core steps: first, it filters for genes with average expression > 0.02 UMI/ cell, and Fano factor > 2; second, it reduces dimensionality to a 50 dim PCA space; third, it constructs a k-nearest-neighbor (kNN) graph in this space; finally, it renders an interactive 2D visualization of this kNN graph using a force directed layout. In this visualization edges of the kNN graph are literally springs that pull together similar cells, while every cell has a magnetic repulsion force that pushes it away from other cells and an intrinsic gravity. The balance of these forces illuminates the topology of how cells are positioned in high-dimension with respect to one another. In this visualization nodes can be moved around to rotate the projection and find the clearest representation; in each new position the graph re-relaxes according to the underlying physics of the force directed layout. The specific steps we took to generate Fig. 4.3A are as follows. Note that a small manual correction was made to the position of the ESC and LMN populations to reduce white space, but did not distort relationships between populations.

1. Load the cells along each differentiation path into SPRING separately.

2. Remove doublets. Doublets were identified by three criteria. 1) they are rare, in line with our experimental expectation for cell doublets, 2) they formed long range connections between large cell groups
in the same timepoint and sample, 3) they do not possess any unique marker genes; all genes they express are a linear combination of two other cell states. We identified approximately 80 doublets in the standard protocol (<3% of all cells), and approximately 40 in the direct programming approach (<2% of all cells).

3. Load the filtered cells from both protocols into SPRING together. To make plots the coordinates from the SPRING representation were exported into MATLAB; cells were either colored by cell state (Fig. 4.3A), or gene expression (Fig. 4.3B).

**Comparison method II: Pairwise cosine similarity of cell state centroids.** We also asked how the direct programming and standard motor neuron differentiation paths were related to each other by performing a pairwise comparison of cell state centroids. Centroids are the average or center of mass in gene expression space for a collection of cells. Centroids can provide a very accurate estimate of global gene expression that averages over the noise intrinsic to single-cell RNA sequencing at the level of individual cells. By computing the cosine similarity between two cell state centroids we are asking how similar these states are in average gene expression. If the paths do indeed split and then reconverge our expectation was to see similarity between early and late states in both progressions, but not between the intermediate states. We chose to perform these comparisons in a PV-gene space constructed from cells in both protocols to make our comparisons as sensitive as possible. We obtained similar but less sensitive results using all genes above a minimum expression value (not shown). A summary of the steps of this analysis is: 1. Combine all cells from both protocols, extract PV-genes (as described above), and z-score normalize their expression. 2. Calculate the centroid of every cluster from each trajectory in this PV-gene space. 3. Compute the pairwise cosine similarity for all direct programming clusters versus all standard protocol clusters. 4. Visualize: we chose to use a heatmap (Fig. 4.3C).

**Comparison method III: Maximum likelihood assignment of single cells to cell state centroids between trajectories.** Finally, we performed an independent comparison of the differentiation paths that did not depend on our definition of cluster boundaries in the direct programming trajectory. We asked whether any poten-
tially rare individual cells or subpopulations in the direct programming trajectory resemble the intermediate states of the standard trajectory. Because we were now dealing with single cells, not averages of many measurements, care was necessary in this analysis to remain robust relative to the noise in single cell measurements. We therefore took a Bayesian approach and reasoned as follows. In our data, each cell is a vector of counts, with one element for every gene. These counts can be viewed as a multinomial sample from some underlying distribution of gene expression. Since each state of the standard differentiation trajectory is defined by a particular gene expression distribution, we can ask: what is the probability a given direct programming protocol cell was sampled from each standard differentiation trajectory cluster? In this usage, probability amounts to a measure of similarity, with high probability indicating high similarity. We obtained similar results working in either a PV-gene expression space, or considering all genes expressed above a minimum counts threshold. The results of this analysis using PV-genes are presented in Fig. 4.3D. We also obtained similar (but more noisy) results using cosine similarity as an alternative distance metric (not shown). The specific steps of our computation are as follows: 1. Extract a set of genes with which to make comparisons (either PV-genes or all genes above a minimum expression threshold). 2. For each cluster of the standard trajectory, calculate the probability of observing a given gene (i.e. the fraction of counts in that cluster from the gene). For genes that are not detected add 1e-07 total counts. 3. For each cell in the direct programming trajectory, calculate the log-likelihood that it was drawn from each of these clusters. This log-likelihood is from the multinomial distribution function using the probabilities obtained in step 2. 4. Identify and tally the maximum likelihood assignments of all direct programming cells. Normalize raw assignments so that they sum to 100 (giving the percentage). Plot the percentage of direct programming cells assigned to each standard protocol state.

Cell cycle gene expression analysis

Cell-cycle activity can be estimated from a cell cycle associated gene expression signature; populations that express higher average levels of cell cycle genes are most likely cycling at higher frequency than a population
with lower level expression of these genes. In Fig. 4.4b and Supp. Fig. A.18 we performed an analysis in this spirit to determine which parts of the DP and SP differentiation trajectories appeared to be proliferative, and to estimate where cells are exiting the cell cycle. We computed a proliferation score that was the aggregate expression of a panel of 21 cell cycle related genes: Aurka, Top2a, Ccn2a, Ccnd1, Ccnd2, Ccnd3, Ccne1, Ccne2, Ccnb1, Cdk4, Cdk6, Cdk2, Cdk1, Cdkn2b, Cdkn2a, Cdkn2c, Cdkn2d, Cdkn1a, Cdkn1b, Cdkn1c, Mcm6, Cdc20, Plk1, and Pcn1. We also computed a cell cycle exit score on the basis of the aggregate expression of a panel of 4 tumor suppressor genes that inhibit the cell cycle: Cdkn1c, Cdkn1b, Cdkn1a, and Cdkn2d. In Supp. Fig. A.18 we show the expression of representative individual genes from this score; in general cell cycle genes were correlated with each other in their expression over cells, as were cell cycle exit genes.

**Lifetime estimate for hypothetical Olig2+ intermediate cells in DP**

We were interested to convert a bound on the population expression of Olig2, obtained by qPCR, into a bound on the lifetime of a hypothetical rare subpopulation that expressed appreciable levels of this transcript. To proceed, we noted that in SP, Olig2 is expressed 5-fold higher than Gapdh. By contrast, qPCR indicated that in DP, Olig2 was expressed 106-fold lower than Gapdh. Since we took qPCR measurements on every day during differentiation, we next reasoned that the maximum possible number of these intermediate cells would exist if they were spread uniformly across our timecourse (otherwise there would be a spike in their number, increasing the chances of their detection). This is unrealistic, but sets a conservative upper bound. Since the total differentiation protocol took 11 days, any Olig2+ intermediate that expressed the gene at the levels seen in SP (5 fold higher than Gapdh) must thus exist for less than $11^{days}/5$, or less than 0.2 seconds. To estimate what this lifetime would be for an Olig2+ population that expressed just 1 molecule per cell, we conservatively allowed for Gapdh expression levels of 1000 molecules per cell (smRNA-FISH studies have estimated 200-500 Gapdh copies per cell). The maximum fraction of Olig2+ cells expressing 1 molecule per cell at any timepoint in DP is correspondingly 0.1% ($1000/106 * 100\%$). Our lifetime calculation becomes $11^{days}/103$, or less than 15 minutes. Since the timescales of mRNA production and degradation are typically
on the order of hours, we therefore concluded that an Olig2+ subpopulation must not exist during DP.

Comparison of motor neurons in vitro with primary motor neurons using single-cell data

How does the transcriptional state of motor neurons produced by both protocols compare to that of motor neurons in vivo? To answer this question we leveraged the ability of single-cell RNA sequencing to compare cell states even within populations that are not pure (also see below for functional comparisons of the phenotypes). We performed three analyses.

1. First, we computed the cosine similarity between the centroids of each cell state in both protocols and primary motor neurons (Fig. 4.3B). The specific steps of this analysis were as follows:

   (a) Combine all cell states from both protocols, and from HB9+ E13.5 primary tissue, extract PV-genes (as described above), and z-score normalize their expression.

   (b) Calculate the centroid of every cluster from each trajectory, and from primary motor neurons, in this PV-gene space.

   (c) Compute the cosine similarity for all in vitro populations versus primary motor neurons, and visualize as a bar graph.

2. Second, we compared single cells from the DP and SP terminal states to pMN reference single cells. The goal of this analysis was to detect whether heterogeneity within the DP and SP terminal states correlated with similarity to pMNs. To gain a broad understanding of the relationship between the DP and SP trajectories to our pMN reference, we initially projected pMN single cells into the PCA-space used in the visualization from Fig. 4.3A (Fig. 4.3C-i). The projection was performed by multiplying the pMN counts matrix (after quality filtering cells and z-score normalizing genes) by the PC-loadings matrix that was built on the DP and SP cells as described above. The combined coordinates of all cells in this space were then used as input to SPRING, which builds and visualizes a kNN graph.
To generate a more refined visualization of how the terminal populations were related to each other we next constructed a new PC-space on EMN and LMN DP and SP cells, combined with pMN, following the same steps of SPRING visualization described above (Fig. 4.3C-ii). This differs from the original projection in that the second PC-space was constructed to reflect variation within only the mature populations of cells (EMN, LMN and pMN). It thus has higher sensitivity to detect subtle differences between MN subpopulations, because the PC-space does not contain dimensions that describe unrelated process that occur in the process of becoming a neuron from an mES transcriptional state, and thus only introduce noise into the comparison of terminal states. To quantify how similar individual cells were to the pMN reference we performed nearest neighbor analysis on the distance matrix underlying this second SPRING visualization. For each DP or SP terminal state, we asked of every cell whether it had a pMN cell among its 50 most similar cells, and found the fraction of cells with similarity to pMN. The number of nearest neighbors chosen, in this case 50, sets the size of the region in which we required a pMN cell to fall in order for a cell to be classified as similar to pMN. 50 cells equated to roughly 2% of the total cells in this visualization, and so provided an appropriate and stringent threshold.

3. Third, we performed differential gene expression analysis, comparing the most mature motor neuron states from each protocol with primary HB9+ motor neurons (Fig. 4.5). This analysis was performed as described above.

qPCR

RNA was isolated using Qiagen Rneasy Plus Kit. Purified RNA was then reverse transcribed using Bio-rad's iSCRIPT cDNA synthesis kit. Quantitative PCR was performed using Bio-rad’s SYBR green Supermix on a CFX96 Real-Time PCR system.

Immunostaining
Antibodies used for immunostaining were anti-Tubb3 (Cell Signaling D7iG9; 1:100), anti-Map2 (Sigma M9942; 1:200), anti-VACht (Synaptic Systems 139 103; 1:200), anti-Isl1 (Abcam ab109517; 1:1000), and anti-Hb9 (DSHB 81.5C10; 1:100). Differentiated cells were fixed in 4% paraformaldehyde for 20 minutes and then permeabilized with 0.1% Triton-X for 15 minutes. After primary incubation for 1 hour, samples were labeled with a secondary antibody conjugated to AlexaFluor647. Samples were co-stained with DAPI before imaging on a Nikon Eclipse TE2000-E microscope.

Electrophysiology

All recordings were carried out at room temperature within 6 days of plating the neurons in 35 mm dish. Whole-cell voltage clamp recordings were made with a Multiclamp 700B amplifier (Molecular Devices) and patch pipettes with resistances of 2-3 MΩ. Pipette solution was 135 mM K-Gluconate, 10 mM KCl, 1 mM MgCl₂, 5 mM EGTA, 10 mM HEPES, pH 7.2, adjusted with NaOH. The external solution was 140 mM NaCl, 5 mM KCl, 2 mM CaCl₂, 2 mM MgCl₂, 10 mM HEPES, and 10 mM D-glucose, pH 7.4, adjusted with NaOH. We used gravity perfusion system connected with Perfusion Pencil® with Multi-Barrel Manifold Tip (AutoMate Scientific) to externally apply 0.5 μM tetrodotoxin, 100 μM AMPA, kainate, GABA, or glycine to the cells. Command protocols were generated and data was digitized with a Digidata 1440A A/D interface with pCLAMP10 software.

Co-culture muscle contraction assays

C2C12 myoblasts were grown in 10% FBS+DMEM media and then differentiated into myotubes by incubating in differentiation medium (2% horse serum + DMEM). After myotubes were formed, the neurons were dissociated by trypsinization and reseeded on top of the differentiated muscle to allow contractions to develop. Video of contractions were taken using Metamorph software and manually counted over 5 second intervals. For stopping assay, 300 μM Tubocurarine (Sigma) was added to the media as an acetylcholine competitor. For labeling of acetylcholine receptors, bungarotoxin (Invitrogen) was used after cells were fixed.
with paraformaldehyde. Similarly to C2C12 cells, ES cells over-expressing MyoD were also differentiated to myotubes using differentiation medium and subjected to co-culture with neurons.

**Microarray Analysis**

The mRNA of undifferentiated iNIL ES cells and NIH 3T3 cells (grow in DMEM+10%FBS) were collected and purified by RNA extraction using RNeasy Plus Extraction Kit (Qiagen). Neurons differentiated by either protocol were first sorted by flow cytometry on a BD FACSaria II machine (Beckton Dickinson, USA) to collect Hb9::GFP+ cells, and then were subjected to RNA extraction in a similar fashion. Collected RNA was then amplified and hybridized to Affymetrix GeneChip Mouse Transcriptome Arrays (MTA 1.0). Results were processed by the Children’s Hospital Microarray Core Facility, and were analyzed using Affymetrix’s Transcriptome Analysis Console and Expression Console software.
Science is about knowing; engineering is about doing.

Henry Petroski

Reverse engineering fate choice mechanisms
5.1 Chapter overview

In this chapter, we explore how scRNA-seq profiling of the differentiation of a complex multi-tissue human organoid can: i) reveal the underlying differentiation pathways, and ii) identify mechanisms that can be manipulated to control differentiation outcomes.

In particular, we investigate how the heterogeneous activation of a single TF, GATA6, reproducibly induces a stereotyped mixture of multiple fates in human organoid and monolayer cultures derived from pluripotent stem cells. By using scRNA-seq to map the underlying differentiation hierarchy, we reveal an initial fate choice between epithelial, ventral mesoderm, and hepatic trajectories that is correlated with Gata6 expression level - suggesting a dose-dependent fate choice mechanism. To test this mechanism we experimentally manipulate Gata6 expression levels and show that it is possible to bias the differentiation of the organoid into desired mixtures of cell states. This work is currently being further developed as a practical or ‘engineering’ application of scRNA-seq based differentiation mapping.


Author contributions: This work was a collaboration of equal contribution between myself and Patrick Fortuna. Patrick performed the stem cell experiments in the lab of Ron Weiss. I performed the scRNA-seq data collection and analysis. We are writing the manuscript together.
5.2 **GATA6 overexpression drives multi-tissue organoid differentiation**

We previously established a human induced pluripotent stem cell (hiPSC) line which harbors a doxycycline-inducible GATA6-transgene at variable copy numbers across the population\(^9\). GATA6-hiPSCs differentiate into a complex multi-tissue culture in response to transgene activation. High GATA6-expressing cells were shown to adopt a hepatic fate, leading to the resulting tissue being described as 'liver-bud'. Here we asked three questions: 1) do GATA6-hiPSCs differentiate differently when cultured in three-dimensions (3D) as organoids; 2) what is the full complement of the cell states that emerge during GATA6-organoid formation; and 3) what mechanism(s) lead to the formation of multiple fates in response to activation of a single transcription factor?

We differentiated the GATA6-hiPSC system in both a monolayer (2D) culture format, similar to that which we reported previously, as well as a new 3D organoid culture system (Fig. 5.1A). Both protocols begin with a five-day transgene activation period in mTeSR\(_{1}\) pluripotency-maintenance media, which initiates differentiation, followed by five-days of further differentiation in minimal APEL\(_2\) culture media (Fig. 5.1B). Three-dimensional organoids are embedded in a 2mm layer of matrigel, while monolayer cultures are adhered to the surface of matrigel coated tissue culture plates.

Overall, the 2D and 3D culture systems differentiated into similar mixtures of cell states but differed in their spatial organization (Fig. 5.1C). Both 2D and 3D GATA6-cultures initially expressed OCT\(_4\) indicating pluripotency. By day 5 post doxycycline (dox) addition, GATA6 was robustly induced across the population, and associated with the activation of marker genes indicating differentiation into mesodermal (HAND\(_1\) - not shown) and endodermal (FOXA2, SOX17) fates. By day 10, these fates matured to show expression of markers CEBPA/ALB, TFAP2A (not shown), and CD93 indicative of further specialization into hepatic endoderm, epithelial, and vasculature fates. NES positive mesenchymal progenitors where also observed (not shown). In the 2D culture system, clear symmetry breaking between SOX17-expressing endodermal and OCT\(_4\)-positive pluripotent was observed by day 5. This self-organization was less clear in 3D where a more
**Fig. 5.1: Gata6-driven differentiation of multi-tissue human organoids.** a) Experimental setup: hiPSCs carrying a dox inducible GATA6 transgene are passaged into either 2D or 3D culture systems for differentiation. b) Experimental timeline: hiPSCs-GATA6 are seeded 1 day prior to Gata6-induction by doxycycline during days 0-5. Cultures are then switched to APEL2 maintenance media for days 5-10. Triangles identify timepoints analyzed by Indrops: blue = control, red = + doxycycline. c) Immunostaining for key cell fate markers in 2D and 3D cultures over time shows complex multi-tissue differentiation. Pluripotent cells (OCT4) transition into endodermal progenitors (SOX17, FOXA2), and eventually pattern into hepatic (CEBPA, ALB) and vasculature (CD93) fates. Scale bars: i-iii and ix-x = 200µm; iv-v = 150µm. vi-viii = 100µm.
uniform organization of FOXA2 was observed. Both systems eventually showed spatial self-organization, including the appearance of CD34 and CD93 vascularized networks by day 10. Spatial organization became comparatively more elaborated in 3D organoids compared to 2D culture over time. 3D organoids showed more highly elaborated vascularization (CD34/CD93) and budded hepatic endoderm (CK19/ALB) structures not seen in 2D.

5.3 Mapping cell differentiation in GATA6-organoids by scRNA-seq

To investigate the hierarchy of cell states generated by GATA6 induction in our 2D and 3D culture systems in detail we performed scRNA-seq over the first ten days of differentiation (Fig. 5.2). We profiled a total of 15,738 cells at days 0, 5, and 10 for both protocols and for a control experiment in which dox was not added to induce GATA6 expression (avg 3.2k UMI/cell). Visualization of these time series data using SPRING revealed a GATA6-dependent branching differentiation trajectory that correlated with chronology (Fig. 5.2A-left). SPRING renders a force directed layout of a kNN-graph over cells based on their gene expression similarity to reveal differentiation progressions. We clustered these data using spectral clustering to reveal major cell states in both protocols and used marker genes specifically expressed by each cluster to annotate the identity of each state (Fig. 5.2B and not shown). The scRNA-seq data indicated clear temporal progressions from pluripotent day 0 cells into each fate present on day 10 of differentiation, involving a series of hierarchical fate choices. We formalized these relationships by measuring the inter-cluster connectivity between cell states across adjacent timepoints (Fig. 5.2A-right).

Differentiation in the 2D and 3D culture systems followed a similar structure of fate choices. During the first 5 days of doxycycline treatment, cells diverged into three clear pathways: i) endoderm progenitors, enriched for markers SOX17, FOXA2, OTX1; ii) ventral mesoderm progenitors, expressing HAND1, FABP7, HAPLN1, PLAT; or iii) ectoderm progenitors, expressing TFAP2A/A2C, SOX2, SOX3, IRX2, DLX5/6 (Fig. 5.2A-B). We observed that the 3D system was biased towards endoderm (51% of cells expressing FOXA2
Fig. 5.2: Mapping Gata6-organoid differentiation by scRNA-seq. a) Left: SPRING plots of scRNA-seq timeseries for each protocol, colored by timepoint (inset) cell state. Right: relationships between cell states over time inferred from graph similarity. Edge weight is proportional to cluster connectivity: thick edges indicate putative differentiation progression. Colors and state names are preserved across panels. b) Expression of key marker genes identifies cell states. -Dox control: NANOG (pluripotency), SOX2 (ectoderm/neural), PAX6 and SIX3 (neural). 2D and 3D systems: FOXA2 (endoderm), SERPINA1 (hepatic endoderm), HAND1 (ventral/cardiac mesoderm), TMEM88 (cardiac mesoderm), CD34 (vasculature progenitors), TFAP2A (ectoderm/epithelial transporter), GABRP (epithelial transporter) and SOX10 (neural crest).
at day 5 compared to 42% in 2D) while fewer cells differentiated into ventral mesoderm (6.3% expressing HAND1 cells in 3D compared to 27% in 2D at day 5) and ectoderm progenitors (1.1% expressing SOX10 cells in 3D compared to 2.6% in 2D at day 5) (Fig. 5.2C). Higher cell density in 3D may favor initial endoderm differentiation. By day 10, each differentiation pathway further matured into one to three distinct cell states. Endoderm progenitors matured into hepatic endoderm (SERPINA1/D1t, APOA1/2, CEBPA, ALB). Ventral mesoderm matured into vasculature progenitors (CD34, FLT1, CDH5, FLI1, SOX18, CD93), cardiac mesoderm (TMEM88, HAND1, H19, BMP4), or erythrocytes (HBA1/B, ALAS2, KLF1, TAL1, GYPA/B). Ectoderm matured into either epithelial transporter cells (TFAP2A, KRT19, GABRP, CDH10, SLC1A6, SLC2A1, SLC2A3, CRH) or neural crest (SOX10, FOXD3, PAX3, TFAP2B). In the absence of GATA6, all cells differentiated towards an anterior neuroepithelial fate, marked by expression of SOX3, SIX3, and PAX6, showing that the observed differentiation hierarchy was dependent on GATA6 expression (Fig. 5.2A). Overall, the proportions of day 10 fates were remarkably constant across the 2D and 3D culture systems, suggesting that the mixture of fates is stable despite likely different local cell-cell signaling in each case. The detailed scRNA-seq data allow investigation of the mechanisms responsible for this stability.

5.4 GATA6 expression level correlates with early fate choice

We aimed to understand the mechanisms driving formation of each distinct fate in our GATA6-driven differentiation system. Since GATA6 expression levels had previously been associated with endoderm formation in this system, we first investigated whether the dose of GATA6 may explain additional fate choices towards mesodermal and epithelial fates. Our scRNA-seq data indicated that this may be the case, revealing that endoderm differentiation was indeed associated with high GATA6 expression at day 5, while mesodermal differentiation was initially associated with intermediate expression of GATA6, and the ectodermal state expressed low or zero GATA6 (Fig. 5.3A-B). Beyond day 5, this relationship became more complicated, as cardiac but not hematopoietic mesoderm descendants upregulated GATA6 to high levels.
Fig. 5.3: Gata6 expression level correlates with output cell fate. a) SPRING plot showing the differentiation progression into the 4 major fates present on day 10. Dashed area indicates the site of the first fate choice into either ectoderm, mesoderm, or endoderm progenitors. b) GATA6 expression during the first 5 days of differentiation. Ectoderm, mesoderm, and endoderm progenitors show low, medium, and high GATA6 expression respectively. Inset: dashed area from a) colored by cell state. c) Jitter plot of GATA6 expression levels in single cells from each of the three types of day 5 progenitor cells. Expression in each cell state is clustered around a distinct median level (dashed lines), with little overlap, consistent with 3 dose-sensitive cell fate thresholds.

To understand whether different GATA6 expression levels could encode distinct fate choice outcomes we examined the distribution of GATA6 expression in each of the three initial fates (Fig. 5.3C). For the fate choice to depend on GATA6-dose alone, the expression level must uniquely predict a specific fate outcome. Indeed we found that expression of GATA6 in each cell state is clustered around a distinct median level, with little overlap, consistent with 3 dose-sensitive cell fate thresholds.

5.5 Manipulating Gata6-dose to tune organoid cell-type mixture

To test whether the three expression levels of GATA6 observed at day 5 were driving differentiation into each fate, we varied the doxycycline concentration to titrate GATA6 levels across the population of differentiating cells (Fig. 5.4-5.5). These results supported the existence of three dose-sensitive GATA6 expression thresholds. We found that high levels of induced GATA6 expression strongly biased differentiation outcomes towards endodermal fates, as shown by enriched bulk gene expression of endodermal markers FOXA2 (Fig. 5.5C), and higher proportions of single cells expressing the endodermal markers FOXA2 by immunostaining (Fig. 5.5A-B). At moderate induction levels of GATA6, endodermal fates were reduced, while mesodermal
fates were enriched: the average expression of CD34 (hematopoietic marker) and TMEM88 (cardiac marker) were enriched (Fig. 5.3C), as was the proportion of cells expressing these markers by immunostaining (Fig. 5.5A-B). When GATA6 induction was very weak, we found few cells adopting mesodermal or endodermal fates, and a strong enrichment of ectodermal fated cells shown by expression of SOX10 (Fig. 5.5). Differentiation into an epithelial fate appeared to require low GATA6 expression, as our control scRNA-seq experiment of cells differentiating without dox induction of GATA6 differentiated primarily into ectodermal anterior neural cells expressing PAX6 and SIX3 (Fig. 5.2).

5.6 Discussion

Understanding how mixtures of cell fates emerge in complex tissue cultures can be a significant technical challenge. We have shown how single cell RNA sequencing can be used to systematically map the differentiation of a multi-tissue human organoid system. We further show how the detailed data can reveal underlying fate choice mechanisms, by revealing the dose-dependent activity of GATA6 during early hIPSC differentiation. Finally, we have begun to leverage this mechanism to engineer specific mixtures of cell states.

We were surprised to find that GATA6-hiPSCs differentiated into a similar mixture of fates whether
Fig. 5.5: Manipulating Gata6-dose to tune organoid cell-type mixture. a) Immunostaining for key lineage markers in day 10 2D cultures show fate outcomes that depend on doxycycline concentration - and therefore GATA6 expression level. At low doxycycline, cultures are enriched for SOX10+ ectoderm cells, and mid doxycycline concentration the culture is enriched for CD34+ mesoderm cells, and at high dox concentration the culture is enriched for FOXA2+ endoderm cells. Scale bars: 200µm. b) Quantification of immunostaining shown in (a). c) qPCR measurements of key lineage marker gene expression further supports dose-dependent fate outputs in GATA6-hiPSCs.
cultured as a 2D monolayer or as a 3D organoid. This was surprising because 3D culture should promote secondary signaling processes that can influence cell fate independently of the genetically enforced control of GATA6 expression. Our results indicate that the GATA6-hiPSC system is primarily governed cell autonomously by GATA6-dose, at least over the first 10 days of culture. This could be further tested by inducing differentiation in isolated cells and in cultures grown under distinct signaling conditions.

The mechanisms by which Gata6 dose encodes distinct fate outcomes requires further investigation. Beyond previous work on this problem\textsuperscript{99}, our scRNA-seq data provides several intriguing clues. We have identified co-factors that associate with the decision to differentiate via each of the three possible initial pathways: ectoderm, mesoderm, or endoderm. We are now testing: i) whether activation of these co-factors is sufficient for differentiation down each pathway, and ii) how each co-factor responds to a specific GATA6 dose. These insights will allow more precise control of cell type mixtures in GATA6-hiPSC organoids and illuminate how dose-dependent activity allows a single transcription factor to encode multiple distinct fate outcomes.

5.7 Materials and methods

Maintenance of hiPSC-GATA6 line

Previously engineered hiPSC line PGP\textsuperscript{1-GATA6-HA (hiPSCG6)}\textsuperscript{99} was maintained in mTeSR\textsubscript{1} media (Stem-Cell Technologies). Cultures were maintained for 5-6 days or until 80% confluency reached on matrigel (Corning) coated 6 well plates. hiPSCG6s were passaged via single cell dissociation with Accutase as per manufacturer’s instructions (Life Technologies) and reseeded at 20% confluency supplemented with 10μM Rock inhibitor Y-27632 (Ri) (StemCell Technologies). Media was changed the following day to remove Ri and maintained thereafter in mTeSR\textsubscript{1} until cultures reached desired confluency for subsequent passaging.

2D differentiation setup

Differentiations were setup by seeding 25,000 dissociated hiPSCG6s single cells / cm\textsuperscript{2} onto matrigel coated 12 or 6 well plates (depending on size of desired differentiation setup) supplemented with 10μM Ri. Media
was changed no later than 12 hours after the addition of Ri to minimize any detrimental effects caused by Ri on hiPSCs. mTeSR1 was supplemented with 1μg/ml doxycycline (dox) to induce GATA6 mediated differentiation unless otherwise stated. Induction was maintained for a total of 5 days, continuing to supplement mTeSR1 with dox. At day 6, cultures were transitioned to APEL2 media (StemCell Technologies) for maintenance and growth and changed daily.

**3D differentiation setup**

For initial aggregation, single cell dissociated hiPSCG6s were seeded at 15,000 cells in mTeSR1 containing 10μM Ri into each well of a 96 well U bottom ultra-low attachment plate (Corning Costar - Sigma) and centrifuged at 300g for 3 minutes. The following day, each well was gently triturated 4-5x times to remove any debris and dead cells attached to the aggregates, followed by aspiration of the media and replacing with mTeSR1 supplemented with 1μg/ml dox to initiate induction of hiPSCG6s. 3D differentiations were maintained for 5 days in mTeSR1 containing dox within the 96 well plates. At day 6, GATA6 organoids were embedded in matrigel (as described in [19]) and transferred to lumox dishes (Sarstedt) and maintained in APEL2 media with daily media changes.

**Immunocytochemistry of 2D Gata6 cell cultures.**

Cultures were fixed in 4% fixation buffer (BioLegend, USA) at room temperature (RT) for 30 min and subsequently washed 3x in phosphate-buffered saline (PBS) (Life Technologies) for 1-2 minutes. Cultures were then blocked and permeabilized in 10% donkey serum (Invitrogen) with 0.1% Triton X-100 (Ajax Finechem) in PBS for 1h at RT. Primary antibodies (see list below) were diluted in blocking solution and incubated overnight at 4°C. Wash 3x with PBX (0.1% Triton X-100 in PBS) for 3-5 minutes. Negative control samples were incubated with 10% donkey serum in PBS without primary antibody. Samples were then incubated at RT for 2 hours with isotype- and species matched alexa-fluor conjugated secondary antibodies together with Hoechst diluted in PBX (as per list below). Wash once in PBX for 2 minutes followed by 2x washes in PBS
for 2 minutes each. Samples that were stored for extended periods of time were mounted in prolong gold antifade solution (Thermo Fisher) all other samples were kept in PBS and imaged subsequently after completion of staining.

**Immunocytochemistry of 3D Gata6 organoid cultures.**

Organoids required for immunocytochemistry were transferred onto transwell polyester membrane cell culture inserts (Sigma-Aldrich) and cultured for a week prior to commencement of fixing and staining – to allow for efficient whole mount staining and imaging. Cultures were then transferred to 24 well glass bottom plates with high performance 1.5 cover glass (Cellvis) and fixed in 4% fixation buffer overnight at 4°C. Samples were washed 3x in PBS for 5-10 minutes, followed by blocking and permeabilization in 10% donkey serum (Invitrogen) with 0.3% Triton X-100 for 2-4 hours (depending on size) at RT. Primary antibodies (see list below) were diluted in blocking solution and incubated overnight at 4°C atop of rotating platform. Negative control samples were incubated with 10% donkey serum in PBS without primary antibody. Wash 5x with PBX for 10 minutes and subsequently incubate samples in isotype- and species matched alexa-fluor conjugated secondary antibodies together with Hoechst (see list below) for 4-6 hours at RT atop of rotating platform. Wash once in PBX for 5 minutes followed by 2x washes in PBS for 5 minutes each. Imaging was carried out on the day of staining in PBS in the 24 well glass bottom plates.

**Primary antibodies.**

2D staining used at 1:300, 3D staining used at 1:200. CD34 (rabbit, Abcam), CD93 (mouse, Abcam), CEBPA (sheep, R&D), FOXA2 (goat, RD), Nestin (NES) (goat, Santa Cruz), OCT4 (rat, RD), SOX17 (goat mouse, RD), GATA6 (goat, Santa Cruz), GATA6 (rabbit, Abcam), Albumin (mouse, Abcam), CK19 (rabbit, Abcam), PDGFRα (rabbit, Cell Signalling) SOX10 (mouse, Santa Cruz). Secondary antibodies were used at 1:500 for 2D and 1:350 for 3D – Rat 488; Sheep 633; Rabbit 488, 546, 633; Mouse 488, 546, 633 and Goat 488, 546, 633. Hoechst-33342 was used at 1:10,000 (Invitrogen).
Imaging and analysis.

Confocal images for both 2D and 3D samples were taken using a Leica TCS SP5 II 405UV confocal microscope (Leica Microsystems). All images were initially processed using LAS X software followed by processing in Fiji (ImageJ). Quantification of immunofluorescent images was processed through a custom algorithm in CellProfiler\textsuperscript{35}. Quantification of vascular networks was processed via Angiotool\textsuperscript{32}.

RNA isolation, cDNA synthesis, and quantitative polymerase chain reaction.

Total RNA was isolated for both 2D and 3D samples with TRI-reagent using the Zymo Direct-zol RNA Miniprep Plus Kit (Zymo Research) as per manufacturer’s instructions. Briefly, for direct lysis of 2D cultures, 100μl of TRI reagent was dispensed per cm\(^2\) of culture surface area. For 3D cultures, 800μl of TRI-reagent was dispensed per 20x organoids (younger cultures), the same volume was dispensed per 10x organoids (older cultures). In both cases, triturate 4-5 times and let the sample sit at RT for 5 minutes before continuing with RNA extraction as per manufacturer’s instructions (Zymo Research). Optional to perform on-column digestion of genomic DNA using DNaseI as cDNA synthesis kit also contains a genomic DNA digestion step. Complementary DNA (cDNA) was synthesized with the iScript gDNA Clear cDNA Synthesis kit (Biorad) as per manufacturer’s instructions. Quantitative polymerase chain reactions (qPCRs) were performed with Kapa Sybr Fast qPCR Master Mix (2x) Kit (KapaBiosystems). qPCR reactions were performed on a LightCycler 96 thermocycler (Roche). All target genes were normalized to endogenous ETFA. Three replicate qPCRs were performed for each transcript and the relative changes in expression of each gene were determined via the 2-\(\Delta\Delta CT\) \(2 - \Delta\) method\textsuperscript{168}.

InDrops single cell RNA sequencing.

\textit{Preparation of single cell suspensions:} Differentiations in 2D and 3D formats were processed for scRNA-seq. For all the cultures and their corresponding time points (D0, D5, D10) to be harvested on the same day, differentiations were staggered such that all samples lined up on the day of harvesting and processing. Harvested samples for scRNA-seq were then processed using the InDrops method\textsuperscript{145,390}. 2D and 3D samples
were first dissociated into single cell suspensions by washing in PBS once followed by incubating samples in Accutase supplemented with 2mM EDTA (Life Technologies) at 37°C for 10-20 minutes, triturating samples 4-5 times every 5 minutes. Samples were then passed through a 20μM filter and centrifuged at 220g for 3 minutes. Sample pellets were then washed in cold 1X PBS/- once, resuspended in 1X PBS/- with 15% Optiprep density media, and kept on ice until processed by InDrops. All samples were processed within 1hr of dissociation to minimize artifactual changes in state.

**InDrops single cell barcoding and sequencing:** We generated barcoded single cell transcriptome libraries using InDrops v3 chemistry as described in Zilionis et al. 2017. In total, we sequenced 15,738 pass-filter cells across all samples. Libraries were indexed and sequenced as a single pool over two Illumina NextSeq 500 high-output runs.

**Single cell data analysis.**

**Read processing:** Raw sequencing reads were demultiplexed and processed into genes versus cells UMI counts matrices using the custom Python pipeline InDrops.py (https://github.com/indrops/indrops). For quantification, reads were mapped against GRCh38 86.

**Filtering for high-quality cells:** High-quality cells were isolated in three steps. First, only cells above a minimum number of UMIs per cell were kept. Second, cells with greater than 20% of reads mapping to transcripts from the mitochondrial genome were excluded; this feature indicates sick or dying cells. Finally, doublets were removed. Doublets were identified as small clusters of cells that lacked specific marker gene expression, typically expressing a combination of marker genes from two other clusters.

**Visualization of cell population structure:** Filtered single cell data was combined for each differentiation condition (control, 2D or 3D), and visualized using SPRING 88. Briefly, data were first preprocessed into a principal component subspace defined on variable genes as described in Klein et al., 2015. SPRING then constructs a kNN-graph in this subspace, and renders a visualization by a force-directed layout.
**Cell state clustering and annotation:** Cell state clusters were defined by applying unnormalized spectral clustering to the kNN-graph underlying the SPRING force-directed layout. We used Ingo Buerk’s MATLAB implementation of Ulrike von Luxurg, 2007 (“A tutorial on spectral clustering”)\(^\text{300}\). We defined the number of clusters, \(k\), such that the data was divided into the largest number of groups that each differed >2-fold in expression of at least 5 genes.

**Inference of differentiation hierarchy:** We identified the most likely progression of cell states by identifying similar gene expression clusters across adjacent timepoints. Our approach was similar to graph-abstractation\(^\text{314}\). The connectivity between clusters across timepoints, shown in Fig. 1C, was defined as the normalized fraction of edges connecting those clusters in the underlying kNN-graph out of all inter-timepoint edges.
Conclusions and outlook
6.1 Chapter overview

In this final chapter we look back and summarize the key contributions of this thesis. We then turn our attention forwards to consider new opportunities, building on this work, to map the mechanisms controlling cell fate and to understand cell type evolution. Elements of the future directions were used for fellowship applications.

6.2 Contributions of this thesis

We have developed new approaches - based on droplet-microfluidic scRNA-seq - to map out cell differentiation using unbiased measurements of populations of single cell transcriptomes. Building these methods required both experimental and computational innovations. Now established, our approaches are quickly becoming a standard tool of developmental and stem cell biology. They provide an extremely efficient method of identifying and characterizing cell states in embryos, adult tissues, and cultured stem cells.

We demonstrated our approaches by analyzing embryos of multiple species and by testing the underlying principles of differentiation of stem cells in culture. A major outcome of these applications was the construction of extremely detailed maps of the gene expression changes associated with the formation of each and every cell type in early vertebrate embryos, which we have made publicly available via online browsers. However, our approaches also allowed several new observations on: cell type evolution, the formation of the neural crest, refinement of multilineage expression at fate choices, the reuse of gene modules across tissues, and the plasticity of stem cells. Finally, we showed how single cell maps of differentiation can guide engineering of stem cell fate in human tissue organoids.

Specific contributions include:

- *Conditions for the dissociation and scRNA-seq of Xenopus embryos*: For embryos, applying scRNA-Seq involves important considerations. The method should not preferentially select a single cell type.
It should be highly efficient in yield of cells. The dissociation procedure adopted should be quick and complete. Background RNA released by lysed cells in the sample should be minimized. These problems are acute for Xenopus embryos, as it has large yolk-filled blastomeres that are sensitive to shear forces and handling. We overcame these problems by developing a method of efficient capture and very little cell lysis. The highlight of the method is a novel dissociation buffer that we formulated based on new principles. We also optimized cell handling procedures for the microfluidic capture of large and fragile cells.

- **Computational methods for differentiation trajectory reconstruction across whole embryos:** Earlier scRNA-seq studies had shown that the data could be used to stratify cell types\(^{158}\), and to reconstruct one dimensional and simple branching differentiation trajectories of cells in culture\(^ {186,174}\). However, the complexity of differentiation in the native setting of the embryo is orders of magnitude higher. Dozens of trajectories emerge simultaneously and there are complex convolutions of temporal, spatial, cell cycle, and differentiation related gene expression signatures that must be disambiguated to understand the relevant dynamics. We developed computational approaches to resolve differentiation trajectories among this complexity across the whole embryo. Our approach relied on nested rounds of clustering to parse population substructure from large spatial gene expression signatures, and employed the use of shared latent spaces across timepoints to accurately measure cell state linkage over time despite strong temporal batch effects arising from clearance of the maternal dowry. The end result was a cell state tree that was an extremely close match to known lineage relationships.

- **Community gene expression resources:** By analyzing >130,000 cells sampled over ten timepoints during the first day of development of the frog embryo, we generated a map that describes the waves of gene expression that gradually restrict the identity of each and every cell type (Fig. 1.2). The map provides an incredibly rich resource of gene expression dynamics in the early embryo. To share the resource we developed a web browser (tinyurl.com/scXen2018) that supports visualization of gene expression
across the entire early embryo, identification of enriched genes in specific states, differential expression analysis between cell states, co-expression analysis of gene pairs, and visualization of the dynamics of gene expression along each cell state tree lineage. We are working with Xenbase to incorporate these features into the centralized Xenopus community resource portal. We have similarly made available the data from the fish to allow evolutionary comparison, and from our studies of stem cell differentiation to allow comparison of distinct differentiation trajectories into motor neurons.

- **Observations on cell type evolution:** By comparing frog and fish single cell data we reached several conclusions about how cell types evolve. First, we found that gene expression similarity does not predict cell type orthology, indicating substantial divergence in gene expression occurs between functionally similar cell states during 435 My of evolution separating them. Second, we identify transcription factors as a the most conserved subset of genes, suggesting a role in maintaining cell identity while the cell phenotype may change. Third, we found that a genes protein sequence (and therefore structure / function) was entirely uncorrelated with gene expression conservation. This indicates that gene function evolves independently of its expression pattern. Finally, we observed that the same cell state can arise by different differentiation trajectories across species, indicating an underlying plasticity in developmental programs. These initial observations provide proof of concept for using scRNA-seq data to study cell type evolution, and motivate deeper comparisons involving more species (see below).

- **Formation of the neural crest:** It has been contentious how the neural crest emerges in vertebrate development. A key challenge has been explaining how the neural crest is pluripotent (able to generate progeny from all three germ layers). There are two possibilities: either the neural crest derives from a non-pluripotent ancestral state in the gastrula, or it arises from a subset of cells that retain pluripotency from the early blastula. The first possibility is more widely accepted, but seemingly involves a form of cell de-differentiation, which is unusual among vertebrate lineages. The second possibility is more simple and was recently supported by evidence of shared pluripotency gene expression between
blastula and neural crest cells\textsuperscript{38}. Our data however showed that there is no evidence at the level of the transcriptome of a special intermediate for the neural crest in gastrula stage embryos, and instead linked the neural crest to the non-pluripotent ectoderm. We therefore weigh heavily on this debate finding strong support for the classical model of neural crest development and against pluripotency retention.

- \textit{Multi-lineage expression at fate choices:} We used our single cell data to survey the frequency with which genes overlap and then refine during fate choices, and to test what types of genes show this behavior. We show that fate choices in the blastula and gastrula involve very high frequency of gene expression overlap in progenitor cells - as many as 40\% of genes that are differentially expressed after a fate choice overlap in the progenitor cells - but that this frequency declines to less that 10\% of genes overlapping at fate choices by 24 hpf. We find that transcription factors are enriched among genes that overlap in progenitor cells before refining, suggesting multi-lineage expression plays a general role in driving cell fate choice in early embryos. This observation builds on previous work that discovered the phenomena of multilineage expression, and provides the first broad support for its general role in the early embryo.

- \textit{Reuse of transcription factors:} We used our single cell data to collect statistics on transcription factor reuse across all lineages in the early embryo, and found that half of expressed transcription factors are reused across multiple tissues within the first day of development, many across more than two tissues. This result underscores the importance of combinatorial gene regulation in vertebrate development. We also show that single cell data can distinguish cases where a reused transcription factor activates the same gene module in multiple tissues, from where it activates a context-dependent gene expression program.

- \textit{Path independence of cell differentiation:} By analyzing the gene expression trajectories induced by two distinct methods to convert embryonic stem cells into motor neurons we show that the same cell state
can be reached by multiple methods and by distinct paths - paths that can have artificial intermediate states. This reveals a special kind of stability of mature cell states, and provides evidence at the single cell level that differentiation is not path-dependent, with implications for both cell reprogramming and evolution. The phenomenon of multiple paths to the same state parallels observations made between the frog and the fish, where we also saw the same cell state on occasion arising by a distinct differentiation trajectory, suggesting that the plasticity we observed in culture reflects plasticity that underlies adaptation during evolution.

- *Engineering organoid differentiation:* Understanding what cell states emerge in stem cell cultures can be challenging. We show how scRNA-seq profiling of multi-tissue human organoid differentiation can: i) systematically reveal the underlying differentiation pathways, and ii) identify mechanisms that can be manipulated to control differentiation outcomes. Our analysis revealed how a single transcription factor induces differentiation into multiple distinct cell fates a dose-dependent mechanism.

6.3 **Outlook**

The single cell transcriptomic approaches developed in this work open myriad possibilities for future research. There are two areas in particular where we envision rapid progress can be made: i) mapping the mechanisms controlling cell fate; and ii) understanding cell type evolution. Ongoing technological and computational innovations in single cell methods will continue to fuel a rapid pace of discovery.

6.3.1 **Towards comprehensive control of cell fate choice**

We have developed new ways to track the detailed molecular events underlying the formation of each and every cell type in developing embryos using high-dimensional single-cell data. Yet ultimately we would like to understand how these complex molecular changes drive cells to decide between alternative fates during differentiation, and how cells stably maintain their identities once chosen. These problems are of both fun-
damental interest and practical value: their answers would enable rational control of cell differentiation in regenerative medicine and a mechanistic understanding of cell de-differentiation in cancer.

The mechanisms by which cells decide between alternative fates and stably adopt particular identities can be tested by perturbing specific factors in cells as they differentiate, and then recording the resulting changes in cell state using single-cell RNA sequencing (scRNA-seq). However, cell fate decisions may result from the combined action of many genes. Ideally, rather than focusing on individual factors at a time, we would be able to comprehensively and combinatorially manipulate and measure genes of interest in single cells. No tools yet exist for this purpose, but recent advances in CRISPR-based gene regulation have brought this goal within reach. In particular, it should be possible to adapt recent loss-of-function screening approaches that combine CRISPR-Cas9 mutagenesis with scRNA-seq readouts, by replacing the Cas9 nuclease with dCas9 fused to a transcriptional activators or repressors (termed “CRISPRa” or “CRISPRr” respectively), and constructing appropriate guide-RNA libraries.

Single-cell CRISPR activation/repression (scCRISPRa/r) screening allows complex gene expression perturbations to be associated with specific changes in genome-wide transcriptional state. This capability has broad potential applications. It would allow tests of at least three specific questions raised by this thesis:

1. How many ways can a motor neuron be programmed? In Chapter 4 we identify two differentiation trajectories from the mESC state into the motor neuron state. However, the two pathways shared an early neural intermediate state. It is therefore currently unclear whether differentiation is entirely path independent, or whether instead the early neural intermediate state is necessary to reach the motor neuron state. To test whether even the early neural intermediate state can be skipped, we need a method to enumerate many possible differentiation pathways and to determine whether all paths necessarily share common intermediates. scCRISPRa would allow this by making it possible to screen for all possible TF combinations that convert mESCs and fibroblasts into MNs, as defined by expression of an Mnx1 reporter gene. The differentiation paths induced by each combination of TFs in this list
of “motor neuron recipes” could then be contrasted by scRNA-seq. This work flow could be adapted to identify reprogramming cocktails for any cell type of interest given a reporter gene for it exists.

2. **How do transcription factors act context-dependently across cell types?** Transcription factors act context dependently when they activate or repress different sets of genes depending on the cell state in which they are activated. This could be measured directly using scCRISPRa, by activating a transcription factor of interest mosaically across the embryo, and identifying gene modules specifically responsive to the TF in each cell state. By using scCRISPRa combinatorially, it would also be possible to understand how transcription factors act in a given cell state conditional on the presence of specific cofactors.

3. **How do cells refine into alternative states during fate choices?** We identified large numbers of genes that overlap and refine during fate choices in early development. Many of these genes may participate in decision making motifs, such as the cross-inhibitory motif that is closely associated with cell state bifurcations. We do not understand whether these circuits: i) cooperate through chaining to perform “proof-reading” of fate choices; ii) are driven by a single gene pair at the top of a hierarchy; or iii) are independent of one another. scCRISPRa would allow this to be tested, by combinatorially perturbing competing regulators in single cells as they make a fate choice and measuring the cell fate outcomes by scRNA-seq.

### 6.3.2 Towards principles of cell type evolution

There are numerous cladograms of gene sequence that confirm the basic paleontological record of vertebrates and reveal the genetic differences that separate species. Since the early 19th century, the similarities and divergences in embryonic and adult anatomy across species have also been extensively characterized. However, in between the genetic and anatomical scales, we have limited knowledge about how cell types and their developmental programs change across species. This gap limits our ability to understand how changes in
genotypes lead to changes in phenotypes across evolution.

The scRNA-seq approaches that we developed provide an unprecedented opportunity to look deeply into these developmental gene expression programs and identify clues about the most consequential evolutionary changes across species. We can now generate detailed molecular maps of differentiation into each and every cell type of extremely diverse organisms with comparatively little labor, cost, and prior knowledge of the animal. The requirements are only that we can dissociate the organism, and that we have access to a reference transcriptome.

Already, by comparing early frog and fish development by scRNA-seq we gained an entirely new window into the evolution of these species. However, this comparison remains limited in important ways. With only two species we cannot properly assess the significance of observing genes that have diverged or maintained their expression profile across tissues. We also cannot assess the dynamics with which cell types change across species, or easily identify the specific events that lead to the evolution of new cell types. To do this we will need to make comparisons of species across a range of evolutionary differences. In particular, the inclusion of comparisons between very closely related species that differ in only a few cell types will be empowering.

Towards this goal, we and several other labs around the world have already started collecting timeseries scRNA-seq data from multiple organisms. We recently collected data from Xenopus laevis, which offers a very near comparison to Xenopus tropicalis (50 My diverged). We also collected timeseries data from the acorn worm (Sarcoglossus kowalevskii) (Fig. 6.1), which is 684 My diverged from X. tropicalis. It is also a near outgroup of the deuterostome subtaxon (which contains vertebrates), which provides an interesting window into the emergence of deuterostome specific characters. Complementing this data, other labs have now collected very similar timeseries data from mouse, C. elegans, and planaria. Additional datasets

---


†Gotteens and Trapnell labs, personal communication

‡Technically not a timeseries dataset, but effectively acts as a timeseries due to asynchronous differentiation in these organisms
Fig. 6.1: Acorn worm scRNA-seq timecourse. SPRING plot of new single cell RNA-seq data sampled across six timepoints during the early development of the acorn worm (Sarcoglossus kowalevskii). Colors indicate timepoint.

(not timeseries) have also been collected from representatives of Porifera (Amphimedon queenslandica), Ctenophora (Mnemiopsis leidyi), Placozoa (Trichoplax adhaerens), Cnidaria (Nematostella vectensis) and Protostomia (Drosophila melanogaster). These can be extended to timeseries simply to allow comparisons of differentiation. Taken together, we are rapidly accumulating better coverage of the tree of life (Fig. 6.2).

As these datasets accumulate, and are annotated in detail, it will be fascinating to dissect the evolutionary changes that have occurred in developmental programs. Two examples of research directions include asking:

1. *How do conserved cell types change in state over time?* Is change gradual, or abrupt? Can we track the gain and loss of specific gene modules as it relates to cell function? Can we identify a subset of genes that is always preserved - i.e. a minimal identity circuit?

2. *What molecular events associated with the emergence of a novel cell state?* By comparing two closely
related species that differ in just a few cell types, one could ask: at what point in the developmental lineage did the new cell state arise? what molecules are associated with the bifurcation? Are they sufficient to drive novel cell state appearance if introduced across species? What genetic changes lead to the change in gene regulation?

These few examples of questions only begin to scratch the surface of what should be possible in this new approach to evo-devo biology.

6.3.3 **Single cell technology development**

Single cell technologies will continue to improve at a rapid pace both experimentally and computationally. Key experimental improvements will be to increase the mRNA capture efficiency of scRNA-seq, and to increase the sample multiplexing capacity to allow systematic testing of many conditions or extremely dense
time-sampling. Efforts have also begun to integrate additional information beyond the transcriptome including chromatin state\textsuperscript{17}, protein composition\textsuperscript{16}, lineage history\textsuperscript{18}, signaling history\textsuperscript{81}, and spatial context\textsuperscript{1,24,0}. Key computational improvements will be to develop manifold learning methods that are sensitive to both local and global structure, and to develop methods to formally infer differentiation dynamics on these manifolds.
Supplemental material
A.1 CHAPTER 1 SUPPLEMENTAL MATERIAL

This section contains:

Supplemental figures A.1 to A.10.
Supp. Fig. A.1: Optimization of dissociation and inDrops procedures for Xenopus. (A) Representative dissociation outcomes using either calcium magnesium free medium (CMFM) as compared to Newport 2.0 dissociation buffer. While CMFM fails to dissociate pigmented animal cap tissue, Newport 2.0 leads to complete dissociation. (B) Representative micrographs of Xenopus blastomeres during microfluidic droplet capture (also see Supplementary movies 1-3). Without optimizing cell handling procedures, lysis of cells during device injection led to severe chip blockages. (C) Cells dissociated using Newport 2.0 are >95% viable for over 2hrs at 22C. Arrows indicate DAPI positive dead cells immediately after dissociation and buffer exchange, and two hours later. (D) Reads per cell barcode histograms; right peak indicates signal coming from cells while left peak indicates background signal from ‘empty’ droplets. With optimized dissociation and handling procedures scRNA-seq runs with <15% total background signal, and <1.5% background signal per cell containing droplet were achieved.
Supp. Fig. A.2: scRNA-seq sensitivity analysis. Cell type abundance (measured as proportion of cells in the embryo) that would allow the detection of N cells with 95% confidence, where N is indicated by line color, as determined by binomial sampling statistics. Values are given as a function of Xenopus stage, accounting for the different numbers of cells were profiled at each timepoint.
Supp. Fig. A.3: Example of nested clustering. To classify cell states, each timepoint was clustered in two rounds, or levels. The top left panel shows a representative level 1 clustering. The three breakout panels show how a nested round of tSNE-DBSCAN clustering on each individual cluster can reveal additional substructure, by focusing on variable gene expression within each subpopulation. Level 1 clustering key: 1 = anterior pre-placodal ectoderm; 2 = somite; 3 = ciliated epidermal progenitor; 4 = goblet cell; 5 = neural crest; 6 = lateral and intermediate mesoderm; 7 = neurons; 8 = lens progenitor; 9 = ventral mesoderm; 10 = neural plate; 11 = eye primordium; 12 = endoderm; 13 = posterior pre-placodal ectoderm; 14 = cement gland primordium; 15 = presomitic mesoderm; 16 = epidermal; 17 = ionocyte; 18 = tail bud; 19 = forebrain progenitor; 20 = intermediate mesoderm - ssg1; 21 = blood; 22 = notochord; 23 = migrating myeloid progenitors; 24 = cranial neuron; 25 = otic placode; 26 = small secretory cells. CNC = cranial neural crest; PNC = posterior (chordal) neural crest.
Supp. Fig. A.4: Computational cross-validation of stage 22 tSNE-DBSCAN clustering by comparison to SPRING-spectral clustering. To assess the robustness of the cell state classifications generated by tSNE-DBSCAN, we reclustered stage 22 scRNAseq data (replicate 1) by an independent method – spectral clustering (see methods) – and compared the two results. Good qualitative agreement between the two approaches was observed: a median of 88.9% of cells in each spectral cluster mapped to its single most similar tSNE-DBSCAN cluster (top), while a median of 79% of cells in each DBSCAN cluster mapped to its most similar spectral cluster (bottom). Occasionally a single tSNE-DBSCAN cluster mapped to multiple spectral clusters (or vice versa), but these disagreements in general did not indicate gross misclassifications. Off-diagonal signal generally involves <20% of cells in a cluster, and reflects subtle differences in cluster boundaries between methods.
Fig. A.5 (following page): Structure of tSNE-DBSCAN clusters is preserved on an independent visualization of the underlying scRNA-seq data. To demonstrate that the gene expression structure underlying our cell state classifications across all time-points was robust to different data processing methods (and therefore robust), we examined the structure of the clustering summarized in Fig. 1B on data visualizations generated by an independent method – SPRING. The plots show that the main text cell state classifications, defined using tSNE-DBSCAN clustering, and indicated by coloring here, retain their clustered structure on independent SPRING visualization of each timepoint. An interactive version of the SPRING plot for each timepoint with full annotations is available online at: tinyurl.com/scXen2018.
Fig. A.5: (continued)
Fig. A.6 (following page): Voting outcomes during ancestor assignment across timepoints. (A) Voting outcomes during mappings between level 1 clusters. The median consensus was >88% across all assignments at all time-points. (B) Representative level 2 mapping of subpopulations within the S18 and S16 epidermis.
Supp. Fig. A.7: Marker gene expression on the cell state tree. Plots show the average expression level per node of the indicated marker genes on the cell state tree from main text Fig. 1.2C. Examples illustrate representative patterns for broad germ layer markers such as sox2, t (brachyury), and sox17.
Supp. Fig. A.8: Comparison of cell state tree structure to alternative cell state coarse graining algorithm. (A) All cell kNN-graph from Fig. 1.2B showing the full Xenopus scRNA-seq timeseries (replicate 1; see Table S1) colored by timepoint. kNN graph was built using the algorithm presented in our sister paper (Wagner et al.). (B) All cell kNN-graph indicating cell state cluster identities. (C-D) Edge assignments between states across timepoints generated by coarse graining of the underlying kNN-graph (C) closely resemble those generated by the cell state tree algorithm (D).
Supp. Fig. A.9: Differentiation asynchrony in the somites. (A) Cell state tree from Fig. 1.2C, colored to indicate the presomitic mesoderm and somitic muscle, which form parallel branches due to differentiation asynchrony. (B) Visualization of single cells (replicate 1) belonging to both branches reveals the underlying asynchronous differentiation process. Cells are colored either by timepoint (left), or by expression of key differentiation genes (right). Cells organize by differentiation progress with immature presomitic mesoderm states on the left (t, msn1, hoxc10), which then differentiate into more mature somitic tissue (actc1). Mixing of early (S14) and late (S22) cells indicates asynchrony. (C) Real time versus pseudotime gene expression dynamics in somitic muscle differentiation. Asynchrony causes early, middle, and late waves of gene expression to blur compared to a pseudotemporal ordering, which respects a strictly ordered gene expression progression.
Supp. Fig. A.10: Expression at stage 11 of individual genes proposed to constitute a retained blastula pluripotency circuit in neural crest cells. This figure shows the individual genes comprising the 'pluri-circuit' score plotted in main text Fig. 1.4E. Each gene is expressed broadly, including in non-pluripotent tissues. The red circled region indicates the boundary between neuroectoderm and non-neural ectoderm cells, that associates with neural plate border by S12, and neural crest by S13. This region shows no special enrichment of these factors. Foxd3 is lowly expressed by scRNA-seq but detected in mesoderm and dorsal endoderm as well as ectoderm by in situ (Xenbase), showing that it is not a special case.
A.2  Chapter 2 Supplemental Material

This section contains:

Supplemental figures A.11 to A.13.

Supp. Fig. A.11: Proportion of cell states and total cells matched between Xenopus and Zebrafish cell state trees. Proportion of matched states (left) or matched cells (right) from the orthologous cell state mapping in Fig. 2.2A.
Supp. Fig. A.12: Changes in germ layer proportions between Xenopus and Zebrafish. Of sequenced cells, zebrafish are 60% neuroectoderm by 24hpf, as compared to 31% in stage 22 Xenopus embryos. Xenopus embryos by contrast have >2-fold more non-neural ectoderm cells (23% compared to 9%) at the same stage.
Supp. Fig. A.13: Global gene expression similarity of matched Xenopus and Zebrafish cell states. Heatmaps show the Pearson correlation between matched cell states (replicate 1) for the indicated species-comparison and pre-filtering stringency. Genes that are predictive of cell state within one species tend to be poorly predictive across species. Self-correlation is computed by jackknifing the data within a species (see Materials and Methods).
A.3 Chapter 4 supplemental material

This section contains:

Supplemental figures A.14 to A.19.
Fig. A.14 (following page): tSNE visualization of marker genes for DP subpopulations. This figure relates to main text Figure 4.2. It shows the raw expression of sample marker genes that were used to identify each subpopulation. Two marker genes are shown per state.
Fig. A.15 (following page): tSNE visualization of marker genes for SP subpopulations. This figure relates to main text Figure 4.2. It shows the raw expression of sample marker genes that were used to identify each subpopulation. Two marker genes are shown per state.
Supp. Fig. A.16: qPCR validation that MN gene expression in DP follows the expected dynamics and does not involve Olig2 induction. a) Measurement of the bulk population gene expression over time helps to confirm that the trajectory inferred from single-cell analysis matches the true dynamical events of our system. These results confirm that key MN differentiation markers are induced on a real timescale that matches their ordering in our inferred trajectory. Terminal genes such as Mnx1 and Tubb3 are upregulated immediately after induction of early progenitor genes such as Sox1. The values plotted are shown in units of fold change relative to their expression in day 0 mESC cultures. b) Plot shows that Olig2 is expressed $\geq 10^6$ times lower than Gapdh during DP. Red line indicates the expected ratio if 0.1% of the population expressed 1 molecule per cell; the observed ratios are below even this low bound.
Supp. Fig. A.17: Joint visualization of DP and SP differentiation paths by tSNE. Like SPRING, tSNE shows similarity between early and late states, but it is ambiguous with respect to the continuous global topology of either path. Between DP and SP, tSNE confirms that ESC, NP, and LMN states between DP and SP are similar. Moreover, SP intermediate states (PNP, PVNP, and MNP) do not overlap with the DP intermediate state (EMN). However, since tSNE emphasizes local similarity relationships the overall joint topology, as revealed by SPRING in Fig. 4.3A, is in this case ambiguous.
Supp. Fig. A.18: Cell-cycle-associated gene expression decreases as cells enter the DP and SP intermediate states and then progressively mature. This figure relates to main text Figure 4.4B. a) Cell cycle gene expression score overlayed on SPRING trajectory shown in main text. The score is the summed expression of a panel of 21 cell-cycle related genes. Cell cycle expression is high in ES cells and neural progenitors from both protocols, but drops sharply upon entry into the transitional states preceding terminal differentiation. b) Shows the average proliferation score along both DP and SP trajectories computed by state (and max normalized). Arrows indicate the same transition as the arrows shown above in panel a. c-d) Example raw expression data for representative individual cell cycle or tumor suppressor / cell-cycle exit genes. Some genes appear to be enriched in specific states, such as Ccnd3 in ES cells and Ccnd2 in neural progenitors from the SP. Expression in cells from each sample are colored using either a blue (ESC), red (DP), or green (SP) colormap.
Supp. Fig. A.19: Heterogeneity within the Mnx1+ E13.5 primary motor neuron population. This figure shows the raw expression of a set of marker genes that were used to identify the subpopulations indicated in main text Figure 4.5A. Two markers are shown per subpopulation, with their expression intensity overlaid on the original tSNE embedding.
A review of IncRNA mechanisms in the nervous system

This appendix includes a manuscript written during G1 year of my PhD, which reviews the ways in which long non-coding RNAs influence the development, function, and evolution of mammalian nervous systems.

Author contributions: This manuscript is of my own conception in structure and content. It is developed from my honors undergraduate research thesis introduction. Accordingly I wrote the initial version of the text. All co-authors contributed to the editing and refinement of the manuscript. Guy Barry prepared the figures.

B.1 Abstract

Only relatively recently has it become clear that mammalian genomes encode tens of thousands of long non-coding RNAs (lncRNAs). A striking 40% of these are expressed specifically in the brain, where they show precisely regulated temporal and spatial expression patterns. This begs the question – what is the functional role of these many lncRNA transcripts in the brain? Here we canvass a growing number of mechanistic studies that have elucidated central roles for lncRNAs in the regulation of nervous system development and function. We also survey studies indicating that neurological and psychiatric disorders may ensue when these mechanisms break down. Finally, we synthesize these insights with evidence from comparative genomics to argue that lncRNAs may have played important roles in brain evolution, by virtue of their abundant sequence innovation in mammals, and plausible mechanistic connections to the adaptive processes that occurred recently in the primate and human lineages.

B.2 Introduction

Advances in genome sequencing technologies during the last decade have enabled an unprecedented scale of transcript discovery. One of the key results has been the finding that non-protein-coding transcripts dominate the transcriptional output of mammalian genomes. It is now generally appreciated that at least 80%
of the human genome is dynamically transcribed across the developing and adult body\textsuperscript{69}, to produce a wide range of small (sncRNAs; <200bp) and long non-coding RNAs (lncRNAs; >200bp). Functional studies have elucidated a wide diversity of miRNA-mediated mechanisms, which influence almost every aspect of metazoan biology (reviewed in\textsuperscript{10,72}). More recently, lncRNAs have begun to receive similar attention from experimental biologists, resulting in the discovery of a variety of lncRNA regulatory mechanisms with similarly pervasive influence. Here we survey these discoveries as they pertain to the development, plasticity, disease, and evolution of the mammalian brain. We emphasize the abundant opportunities for discovery represented by the wealth of as yet uncharacterized nervous system lncRNAs.

B.3 LncRNAs are abundant and precisely regulated in mammalian nervous systems

The GENCODE\textsuperscript{105} and NONCODE\textsuperscript{108} consortia have annotated 10,000 – 50,000 lncRNA genes in the human genome to date. Many are located in regions historically termed ‘gene deserts’, between protein-coding genes, while others overlap protein-coding genes in both antisense and sense orientations. They range from small single-exon loci to large multi-exonic transcripts with several alternative splice forms. Remarkably, 40\% (4,000–20,000) of these are expressed specifically in the brain\textsuperscript{64}. This number is strikingly large when compared to current predictions that the human genome contains 20,000 – 25,000 protein-coding genes\textsuperscript{109}, and around 2,500 miRNAs\textsuperscript{148}, with many fewer that are specific to the nervous system. It begs the question – what (if any) is the functional role of these many lncRNA transcripts in the brain?

Arguing that widespread functional roles may exist for nervous system lncRNAs, their expression is dynamically regulated during development\textsuperscript{5,17,162,199,216}, and in response to neuronal activity\textsuperscript{11,199,216}. It is also often highly restricted to specific brain regions in adult mice, such as the hippocampus or particular cortical domains\textsuperscript{189}. In fact, it has recently been determined that lncRNAs provide more information about cell-type identity during mammalian cortical development than protein-coding genes\textsuperscript{197}. These dynamics and region-specific expression patterns are coordinated by cell-type specific or activity-dependent transcription factors,
and canonical changes in chromatin state at lncRNA loci\textsuperscript{[139,226].}

While some lncRNAs, such as enhancer RNAs, or antisense RNAs (Box 1), may have tissue-specific expression as an indirect consequence of some other mechanism, most lincRNAs (>7,000 in the human genome) have their own independently regulated promoters that undergo canonical transcription factor binding and chromatin remodeling events independent of any other known functions performed at that locus\textsuperscript{[40,96,226].} Obviously, regulated expression does not prove a functional role for lncRNAs, but it is consistent with it. The evidence for lncRNA functionality in the nervous system is developed below as the main subject of this review.

\subsection*{B.4 Evolutionary conservation of lncRNA loci}

If functional, we might naturally expect lncRNA loci to show evidence of sequence conservation. Indeed, lncRNAs possess highly conserved promoters whose transcription factor binding sites correlate with their tissue-specific expression patterns\textsuperscript{[64,96].} They also have highly conserved splice-junction motifs\textsuperscript{[21]}\textsuperscript{[21].} In contrast, lncRNA gene bodies show relatively low evolutionary conservation, similar to that observed in many cis-regulatory sequences and ancient retrotransposons that have been maintained in the mammalian lineages (which may or may not be functional). This observation prompted some to argue that lncRNAs do not have important biological functions, without much consideration of the circularity of ‘conservation’ indices and the likelihood that regulatory sequences not only have different structure-function constraints but are also the major sites of adaptive radiation\textsuperscript{[219].} Moreover, experimental studies have now shown that the function of specific lncRNAs can be preserved despite this apparent lack of primary sequence conservation\textsuperscript{[216,290].} In fact, several human lncRNAs have been shown to phenotypically rescue depletion of their homologs in zebrafish\textsuperscript{[290].} Thus it seems that the low conservation of lncRNA gene bodies relative to protein-coding genes reflects lower sequence constraint, rather than a lack of functional importance, with other studies showing higher conservation of RNA structure\textsuperscript{[262].} Further supporting this view, several studies have now

168
used mutant mouse model systems to understand the role of lncRNAs in the brain. The results of these studies, notably the PANTR1, VISC2, and EVF2 lncRNA loci, indicate that a large fraction are required for mammalian life and development\textsuperscript{241}.

Is lncRNA function truly independent of protein-coding potential? In short, this seems to be the case for the vast majority of lncRNAs. Comparison of RNA sequencing and mass spectrometry data for at least two cell lines indicates that 92% of annotated lncRNAs produce no detectable peptides [\textsuperscript{14,64}, also see\textsuperscript{86,98}].

B.5 Molecular mechanisms of lncRNAs in nervous system development

Nervous system development is a complex and highly stereotyped process that requires precise spatiotemporal regulation of stem/progenitor cell proliferation and differentiation. These developing populations of cells must also form appropriate connections with each other if the brain is to function properly. In this section we introduce emerging mechanistic roles for lncRNAs in controlling both these processes, under separate subheadings (also summarized in Figure 1).

B.5.1 Stem/progenitor cell proliferation and differentiation

Developmental cell-fate choices are made by the sequential activation of cell-type specific gene regulatory programs in proliferating embryonic stem/progenitor cells. LncRNAs control this process at various stages along the progression from pluripotent cells, which are found in the early embryo, through to the terminal cell-types found in the mature mammalian brain (Figure 1).

\textit{Insights in vitro}: The exit from pluripotency and early neural differentiation has been studied extensively using in vitro model systems such as embryonic stem (ES) cells. In mouse ES cells, systematic loss-of-function studies have identified dozens of lncRNAs that are necessary for establishing pluripotency or driving neural lineage entry\textsuperscript{97,216,254}. These genes are regulated by canonical pluripotency transcription factors such as OCT\textsubscript{4}, SOX2 and NANOG. Many in turn exert their regulatory influence by directing transcription factors

169
Fig. B.1: Regulatory mechanisms of lncRNAs.
or chromatin remodeling machineries to specific lineage specifying genes. For example, the lncRNA RMST is itself regulated by REST, which induces its expression during neural differentiation in vitro. RMST then drives the recruitment of the neural transcription factor, SOX2, to key neurogenesis promoting genes, such as DLX1, ASCL1, HEY2, and SPS in trans^{292} (Figure 1). Loss of RMST blocks exit from the ES cell state and initiation of neural differentiation and so this lncRNA is required for neural differentiation.

Another lncRNA, TUNA, regulates neuronal gene expression by a similar mechanism. TUNA forms a complex with three RNA binding proteins, NCL, PTBP1, and hnRNP-K, that together localize to neural gene promoters in differentiating mouse ESCs^{161}. Knockdown of TUNA, or any one of the three interacting RNA binding proteins, is also sufficient to inhibit neural differentiation^{163}. Amazingly, experiments in zebrafish showed that this functional role was conserved across relatively distantly related vertebrates, further reinforcing the deep requirement for lncRNA-driven neural lineage commitment.

TUNA and RMST exemplify how lncRNAs can control cell-fate choices by directing transcription factors and chromatin remodeling machineries to important target loci. There are other examples, including the lncRNA DALI, which drives the expression of an essential neuronal differentiation gene expression program in neuroblastoma cells. Genomic target mapping by Capture Hybridization Analysis of RNA Targets (CHART) revealed that this is mediated not only through direct interactions with the transcription factor POU3F, and the DNA methyltransferase DNMT1, but also with thousands of target loci across the genome in trans^{46}. Similarly, the lncRNA PAUPAR interacts with the PAX6 transcription factor, and localizes to specific promoter loci including SOX2, NANOG, and HES1, to regulate a transcriptional program that influences the cell-cycle profile and differentiation of neuroblastoma cells^{294}. LncRNA-driven transcription factor localization exemplified by TUNA, RMST, DALI and PAUPAR represents an attractive mechanism for regulation of cell-fate choice, because it can allow complex gene expression programs to be controlled by a single lncRNA.

It has been suggested that a large fraction of lncRNAs (i.e., hundreds to thousands of other lncRNA genes) also operate this way during differentiation. This is largely based on the observation that more than
30% of IncRNAs in mouse and human ES cells physically interact with particular chromatin modifying complexes, and have correlated expression with their targets \(^{97,138}\) (also see Box 1). Although attractive and plausible, the high risk of non-specific interactions between nuclear proteins and large RNA molecules necessitates careful validation of the functional relevance of these seeming interactions, to eliminate the possibility of indirect or protein-independent effects misleadingly creating the detected expression correlations.

**Examples in vivo:** Regulation of cell-fate choice by IncRNAs has now also been studied in vivo, where it has been shown that IncRNAs control stem cell turnover and the specification of particular lineages in the embryonic mouse brain. EVF\(_2\) was the first nervous system-specific IncRNA to undergo detailed mechanistic characterization in vivo\(^{26}\). It was shown that genetic deletion of EVF\(_2\) in mice disrupts the excitatory to inhibitory neuron balance in the postnatal hippocampus and dentate gyrus. This imbalance was characterized by reduced synaptic inhibition, caused by failure of GABAergic interneuron specification \(^{26}\). Mechanistically, EVF\(_2\) recruits the transcription factor, DLX, and methyl-CpG binding protein, MECP\(_2\), to regulatory regions controlling the expression of interneuron lineage genes including as DLX\(_5\), DLX\(_6\) and GAD\(_1\), by both cis- and trans-acting scaffolding mechanisms \(^{26}\). EVF\(_2\) has also been implicated in controlling the methylation state of the DLX\(_5/6\) enhancer control region, which regulates expression of the DLX\(_5/6\) locus through a chromosomal looping mechanism \(^{21}\). How this methylation control by EVF\(_2\) is integrated with its recruitment activities of DLX and MECP\(_2\) is unknown. Nonetheless, EVF\(_2\) is a clear example of the developmental significance of IncRNA regulatory mechanisms in vivo. Several further examples are listed below.

The IncRNA PNKY is expressed in the nucleus of dividing neural stem cells (NSCs) in the developing mouse and human brain, where it acts as a regulator of NSC turnover \(^{225}\). Depletion of PNKY leads to loss of the NSC phenotype in ventricular zone cells, and expansion of the transit amplifying neuronal progenitor pool, in postnatal mouse brains. Thus, PNKY controls the balance between self-renewal and neuronal differentiation in dividing NSCs. PNKY appears to execute this function through regulation of an important alternative-splicing pathway that involves an interaction with splicing regulator PTBP\(_1\) \(^{225}\), though precise epistatic relationships in this pathway remain to be resolved.
The intergenic lncRNA, linc-BRN1B, controls differentiation of delaminating neural progenitor cells in vivo. Deletion of the linc-BRN1B (Pantr2) locus results in significant loss of upper cortical layers (II/III-IV), as well as a reduction in barrel number, and size, within the somatosensory cortex, and posteriomedial barrel subfield, of developing mouse pups. These reductions originate with loss of basal cortical progenitors, which subsequently drives precocious migration and differentiation of lower layer neurons. Linc-BRN1B appears to regulate the levels of its neighboring BRN1 protein, which has been suggested to be important for its ability to regulate basal cortical progenitor turnover. Together these studies on EVF2, PNKY, and linc-BRN1B firmly establish the in vivo relevance of IncRNAs in the regulation of nervous system development.

Several other IncRNAs have been associated with cell-fate choice in vivo, in the developing retina, though their precise mechanisms of action are less clear. In fact some of the first IncRNAs to be implicated in the control of cell-fate choice in the nervous system were identified by early studies in the retina. TUG1, a nuclear restricted lncRNA, and VAX2OS1 have been implicated in control of retinal cell-type specification and proliferation. SIX3OS controls the specification of photoreceptors, bipolar cells, and Muller glia, possibly through regulation of SIX3 target genes via a mechanism that involves interactions with EZH2 and EYA proteins. GOMAFU is expressed in the nucleus of dividing neural stem cells and differentiating neurons. It regulates splicing of several neuronal genes including DISC1, ERBB4, and WNT7B, likely through interactions with splicing proteins SF1, SRSF1, and QKI. Depletion of GOMAFU in embryonic mice leads to increased amacrine cell and Muller glia differentiation, as well as an improperly regulated transition of actively dividing ventricular zone progenitor cells into differentiating neurons as they migrate outward into the cortical plate.

Overall, IncRNAs regulate cell-fate choice and stem/progenitor cell turnover during neural development in vitro and in vivo. To do this they execute lineage-specific gene expression programs by organizing epigenetic, transcriptional, or post-transcriptional protein machineries in space and time, in response to differentiation signals (Figure 1, also see Box 1).
B.5.2 Neurite elaboration and synaptogenesis

As the nervous system develops, an intricate process of neurite elaboration unfolds, such that populations of neurons faithfully establish the connections required for normal brain function. This is an enormous regulatory task. Examples of how lncRNAs contribute are beginning to emerge, and are introduced here.

The first example of a lncRNA that regulates synaptogenesis was BC1/BC200, which was among the first lncRNAs to be ascribed a concrete molecular mechanism in any tissue. BC1/BC200 is expressed in the developing and adult nervous system where it is actively trafficked to dendrites\textsuperscript{204}. There, it interacts with FMRP and translational machineries including eIF4a and poly(A)-binding protein (PABP), to control 48S complex formation, and repress local translation in synapses\textsuperscript{103,123}. Through this mechanism, BC1/BC200 regulates spatially restricted synaptic turnover in vivo\textsuperscript{59,260,287}. Some of the consequences of this regulation are known, and are discussed in the next section under behavioral phenotypes associated with lncRNA regulation of neural plasticity.

More recently, it was found that several important proteins that control neurite elaboration, including BDNF, GDNF, and EphB2, are regulated by overlapping antisense lncRNAs in cis\textsuperscript{194}. For example, inhibition of the lncRNA BDNF-AS, which is transcribed antisense to the BDNF growth factor gene, resulted in a 2- to 7-fold increase in BDNF protein levels, which was further associated with reduced EZH2 recruitment and altered chromatin state at the BDNF locus. Elevated BDNF expression then drove neuronal outgrowth, differentiation, survival, and proliferation, both in vitro and in vivo\textsuperscript{194}. Thus, antisense lncRNAs control developmental neurite elaboration through regulating local gene expression, potentially via interactions with transcription factors and chromatin remodeling machinery in cis. Antisense lncRNA transcripts are remarkably pervasive in mammalian genomes\textsuperscript{77,131}. Examples such as BDNF-AS may thus prelude a form of lncRNA regulation that is of widespread developmental influence. Another exciting feature of the work by Modaressi et al. was their demonstration that antisense lncRNA expression could be easily modulated in vivo by antisense oligonucleotides, to specifically induce expression of their overlapping protein-coding gene,
opening the possibility of lncRNA targeting therapeutics that activate gene expression\textsuperscript{194}.

The lncRNA MALAT\textsubscript{i} represents an interesting case study in recent lncRNA research, due to conflicting in vitro and in vivo results regarding its role in neurite elaboration. MALAT\textsubscript{i} is abundantly expressed in neurons and is enriched in nuclear speckles in a transcription-dependent manner. In vitro, in cultured hippocampal neurons, MALAT\textsubscript{i} has been shown to actively recruit SR-family splicing proteins to transcription sites to control the expression of synaptogenesis-related genes\textsuperscript{22}. Moreover, knockdown of MALAT\textsubscript{i} in this system results in a decrease in synaptic density, while overexpression reciprocally increases synaptic density\textsuperscript{22}. Thus MALAT\textsubscript{i} appeared to regulate synaptogenesis by modulating synapse formation/maintenance genes via a splicing-protein-enriched nuclear-domain. However, it was subsequently shown that MALAT\textsubscript{i} knockout mice surprisingly showed no overt phenotype, either in behavior or in neuropil density\textsuperscript{124}.

There has therefore been significant concern about the correspondence of lncRNA behaviors in vitro to their developmental role in vivo. It is possible that closer examination of the neurological and behavioral phenotypes of MALAT\textsubscript{i} knockout mice could reveal more subtle phenotypic defects, or that special redundancies in vivo could potentially compensate for loss of MALAT\textsubscript{i} function. But nonetheless, such results underscore the importance of completing lncRNA functional analyses with thorough in vivo characterization, although negative results are inconclusive. We note that most of the lncRNAs discussed by name in this review have been characterized in vivo and so meet this more stringent criteria.

\textbf{B.6 Molecular mechanisms of lncRNAs in neuronal plasticity}

Neurons are able to change their set of synaptic connections and the relative strength of each of these connections over time in response to sensory experience and other environmental cues. This so-called plasticity underlies learning, memory, and cognition, as well as the brain’s ability to recover from injury or insult. Many molecular pathways that implement plasticity are known [for reviews see\textsuperscript{22,31}]. However, these descriptions are still incomplete. Here we introduce studies that have begun to implicate lncRNAs in the control of neu-
Emerging functional studies demonstrate that lncRNAs play a wide role in neuronal development from early neural differentiation (left) to late stage synaptogenesis (right). LncRNAs are necessary for early neurogenic commitment, during which they recruit transcriptional machineries to specific neural gene promoters (a-d). In dividing neural precursors, lncRNAs also control later cell-fate choices, by similarly scaffolding epigenetic machineries, transcription factors, or splicing regulators, around specific target-gene loci. Examples shown are interneuron specification (f) or retinal cell development (g, h). Late developmental processes, such as neurite outgrowth and synapse maturation (i), as well as regulated cell death in neural circuit establishment (j), are regulated by natural antisense lncRNAs. Additionally, lncRNAs regulate synapse function through their ability to (k) recruit splicing factors to relevant transcripts and (l) through the recruitment of translational repression machineries to specific target mRNAs within local synaptic environments.
ronal plasticity (summarized in Figure 2). Already two of these lncRNAs, KCN2 AS and BC1/200, have been convincingly tied to concrete behavioral phenotypes. Many others remain exciting candidates for future exploration.

B.6.1 **Transcriptional regulation in response to injury and neuronal activity**

Changes in expression of genes such as ion-channel components, or signaling proteins, can dramatically alter the excitability and functional properties of a neuron. For example, potassium channels mediate potassium ion influx during neuronal action potential propagation. The shape of an action potential and the sensitivity of a potassium channel to depolarization can be tuned by modifying the stoichiometry of potassium channel protein subunits. Thus the regulation of these components in response to environmental cues can contribute to neuronal plasticity.

KCN2 is a core potassium channel subunit, whose expression is regulated by an overlapping antisense RNA in response to peripheral nerve injury and in neuropathic pain. In a healthy rat model system KCN2-AS is expressed only lowly, in a subset of 20% of dorsal root ganglion (DRG) neurons, while KCN2 is expressed highly in most DRG neurons. However, in response to peripheral nerve injury, KCN2-AS is strongly induced by the MZF1 zinc-finger transcription factor across the DRG neuronal population independently of KCN2. Elevated KCN2-AS then selectively down-regulates KCN2 mRNA and protein both in vitro and in vivo. The precise mechanism through which KCN2-AS down-regulates KCN2 mRNA and protein awaits future elucidation, though it may involve direct binding between sense and antisense transcripts given their significant overlap, or competition for some DNA or RNA binding factor that ordinarily drives KCN2 expression. In any case, this regulation alters the functional properties of DRG neurons in living rats. Specifically, overexpression of KCN2-AS in the DRG was shown to reduce total voltage-gated potassium current and increase neuronal excitability, producing mechanical and pain hypersensitivities that are core neuropathic pain symptoms in the clinic. Remarkably, blocking KCN2-AS induction attenuated these symptoms following peripheral nerve injury. KCN2-AS thus modulates neuronal plasticity in
response to peripheral nerve injury, and represents a potential therapeutic target in the treatment of human neuropathic pain.

There is some evidence that other antisense RNAs also contribute to the regulation of neuronal plasticity through controlling signaling molecule expression, but in response to neuronal activity. BDNF is an important growth factor in the nervous system that controls the decision between synaptic maintenance and elimination in response to sustained versus sparse activity respectively. This regulation underlies the synaptic consolidation hypothesis. As introduced in the developmental mechanisms section above, it was recently shown that BDNF expression is regulated by an overlapping antisense-RNA, BDNF-AS. Interestingly, BDNF-AS expression is activity-dependent. It may thus play a mechanistic role in coupling neuronal activity to BDNF expression and synaptic turnover in neuronal plasticity, though this awaits direct empirical support.

Generally speaking, regulated transcription in response to neuronal activity is a central process in long-term neuronal plasticity (reviewed in ). Such activity-dependent transcription links the transcriptional output and thus protein composition of a neuron to its recent firing history, and is required for canonical Hebbian learning, for example. In a screen for such activity-dependent transcripts, Kim et al. identified thousands of enhancer-associated IncRNAs (eRNAs) that were rapidly induced by >2-fold following depolarization of mouse cortical neurons by potassium chloride in vitro. Though the majority of these transcripts currently have not been subjected to careful functional interrogation, their expression changes correlate strongly with changes in expression of nearby protein-coding genes. In other biological systems an increasing number of mechanistic studies have shown that eRNAs are essential for enhancer function. eRNAs drive recruitment of the mediator complex, transcription factors such as CBP, NPAS, CREB, and SRF, and RNA polymerase II, to enhancer loci, and are required for their transcriptional promoting activity at target genes. There is no reason a priori to suspect that this does not also occur in the nervous system. We suggest that mechanistic exploration of activity-dependent eRNAs in neuronal plasticity is an exciting topic for future exploration.
Similarly to Kim et al., others have shown that hundreds of IncRNAs that are not eRNAs are also dynamically regulated by neuronal depolarization in vitro\textsuperscript{16,166}. Among these, GOMA\textsubscript{FU} and MALAT\textsubscript{1} are potentially interesting examples. Both are abundantly expressed in neurons, and form ribonucleoprotein complexes within the nucleus that are enriched in splicing proteins. Though speculative, such transcripts may couple neuronal activity to specific post-transcriptional modifications in neuronal plasticity.

### B.6.2 Translational regulation at synapses

Translational control of neuronal plasticity plays a key role in regulating long-term changes in neural circuits underlying learning, memory and behavior (reviewed in\textsuperscript{15}.

The BCI/200 IncRNA regulates translation of specific mRNAs within local synaptic domains in response to neuronal activity. The mechanism through which BCI/200 controls synaptic translation was introduced in the developmental mechanisms section above. Briefly, the BCI/200 IncRNA is actively trafficked to neuronal dendrites, where it acts as a scaffold that interacts with FMRP and translational machineries including eIF\textsubscript{4A} and poly(A)-binding protein, to control 48S complex formation and repress local translation in synapses (see above). Beyond this previous discussion, here we note that BCI/200 expression is in fact dynamically upregulated at specific synapses by local neuronal activity\textsuperscript{203}. Thus, BCI/200 represents a canonical plasticity gene that modifies the protein composition of synapses in response to neuronal activity. Increased neuronal activity in a particular dendritic region would cause higher local expression of BCI/200, which would then negatively feedback on local translation rates.

Genetic deletion of BCI/BC200 in mice results in uncontrolled group I metabotropic glutamate receptor stimulated synaptic translation, neuronal hyperexcitability, convulsive seizures, anxiety, and exploratory behavior defects\textsuperscript{159,260,327}. Thus, it is clear that translational control by BCI/200 is an essential neuronal plasticity mechanism; its breakdown leads to abnormal neuronal activity and diverse behavioral defects. This example extends the known influence of IncRNAs in the nervous system outside of the nucleus, and into the synaptic environment. It is to our knowledge the only example to date of a IncRNA that regulates translat-
tion in neurons.

B.7 Implicating lncRNAs in neural disease pathogenesis

Since lncRNAs regulate nervous system development and function, it makes sense that their dysregulation or mutation would cause neurological disorders. Indeed, genome-wide association studies (GWAS) and comparative transcriptomic studies have associated lncRNAs with conditions including schizophrenia, Alzheimer’s disease, autism spectrum disorder (ASD), Asperger’s syndrome, amyotrophic lateral sclerosis (ALS), bipolar disorder, depression, Parkinson’s disease, neurofibromatosis, neuropathic pain, attention deficit hyperactivity disorder, epilepsy, brain cancers, and a range of cognitive performance metrics.\(^1\) Although the majority of this work has provided association without yet clearly establishing causation, several examples of causal lncRNA disease mechanisms are beginning to emerge. We introduce some of these below. For a broader coverage we refer the interested reader to two comprehensive recent reviews on the topic\(^2\).

B.7.1 Autism spectrum disorder (ASD)

ASD is a neurodevelopmental condition of complex etiology that is characterized by social and cognitive impairments and repetitive behaviors\(^5\).\(^6\).\(^7\).

GWAS on ASD have identified strongly disease-associated variants in the chromosomal region 5p14.\(^1\). The lack of protein-coding genes in this region made this association initially difficult to explain. Kerin et al. later found that these variants reside near a 4kb lncRNA that is transcribed antisense to moesin pseudogene 1 (MSNP1AS), and that shares 94% sequence identity with the 4kb mature MSN mRNA\(^13\). This finding was immediately interesting because MSN is itself a well-known regulator of synapse development and function\(^8\).\(^9\).\(^10\).\(^11\).\(^12\). Thus, if MSNP1AS locus variants altered the regulation of MSN protein expression or function, it could explain how these variants might underlie the development of synaptic dysfunction in
Fig. B.3: Molecular mechanisms of IncRNAs in synaptic plasticity. LncRNAs respond to neuronal activity or injury and modulate synaptic properties (top) or gene expression (bottom) in neuronal plasticity. (a) BDNF-AS is the natural antisense transcript to BDNF, itself a key contributor to synaptic function. By dynamically repressing BDNF expression in response to neuronal depolarization, BDNF-AS modulates synaptic function. (b) MALAT1 expression levels control spine maturation and synapse formation, both during development, and in response to neural activity, by recruiting splicing factors into a currently poorly understood splicing domain within the nucleus. (c) BC1/200 expression is modulated by neuronal activity, and controls the translational repression of specific target mRNAs within synapses, through a mechanism involving direct recruitment of translational machineries. (d) eRNAs are emerging as key activity-dependent regulators of synapse development through their recruitment of multiple proteins required for neuronal gene expression to transcription start sites. (e) GOMAFU controls the activity-dependent release of splicing factors from a nuclear domain to regulate gene expression patterns and splice variant distributions that influence behavior in mice. (f) In the peripheral nervous system, KCNA2-AS is induced in response to nerve injury. It thus dynamically down-regulates expression of the KCNA2 potassium channel subunit, which alters neuronal firing properties. This mechanism is a key driver of neuropathic pain symptoms in rats.
ASD.

Several pieces of evidence support this hypothesis\textsuperscript{137}. First, all three SNP genotypes in the MSNPtAS locus are significantly associated with the expression level of MSNPtAS. The expression level of MSNPtAS is itself also positively correlated with expression of MSN mRNA and protein in the brain, and both MSNPtAS and MSN are overexpressed in ASD patient brains by $>10$ and 2 fold respectively. Second, experiments in vitro indicate that MSNPtAS can directly bind MSN mRNA through its strong sequence homology. And finally, overexpression of MSNPtAS directly alters the expression of MSN protein levels, though these results were sometimes conflicting between experiments conducted in different cell lines. Though yet to be carefully validated in vivo, on the basis of these results, Kerin et al. suggest that MSNPtAS may regulate MSN protein by binding to and stabilizing MSN mRNA, and that this mechanism may causally connect SNP variants in the MSNPtAS locus to ASD pathogenesis.

\subsection*{B.7.2 Schizophrenia}

Schizophrenia is a psychiatric disorder associated with behavioral abnormalities, cognitive and emotional impairment, and psychosis. It commonly manifests between the ages of 10 and 40, has high heritability (estimated at 80\% in twin studies), hundreds of associated risk loci, strong environmental risk factors, and affects 1\% of the population\textsuperscript{235,270,272}.

No unique set of abnormalities is currently sufficient for diagnosis of schizophrenia. Instead it has been speculated that diverse molecular processes converge in various combinations in the pathophysiology of disease\textsuperscript{104}. Though this pathogenic process is complex, several molecular signatures stand out. For example, changes in the splice-isoform distributions of ERBB\textsubscript{4}, an adhesion molecule and receptor tyrosine kinase\textsuperscript{137}, and DISC\textsubscript{1}, a functionally pleiotropic and broadly interacting intracellular molecule\textsuperscript{29,207}, occur frequently and are thought to directly contribute to pathogenic neural development and function in schizophrenia\textsuperscript{120,291}. In some cases direct mutation of DISC\textsubscript{1} or ERBB\textsubscript{4} is thought to underlie their pathological splicing, but for many other patients the root cause of splicing defects is unknown.
We recently produced evidence that the IncRNA GOMAFU may be involved in driving this aberrant splicing of DISC1 and ERRB4 in schizophrenia\(^5\). GOMAFU is an interesting candidate gene because multiple independent studies have previously associated its mutation or dysregulated expression with schizophrenia risk\(^{\text{54,65,69,126,127}}\). It also forms a ribonucleoprotein complex in the nucleus, which is enriched in three splicing proteins, SRSF1, SF-1 and QKI, providing a plausible mechanism that could be dysregulated to generate schizophrenia splicing phenotypes\(^9\). In healthy mice, GOMAFU controls cell-fate choices in the developing retina, and is also highly expressed in the CA1 region of the hippocampus and in large excitatory projection neurons of the cortex\(^{\text{166}}\). There, it is dynamically regulated by neuronal activity\(^{15}\). However, in contrast to these natural and reversible expression level changes, we found that chronic downregulation of GOMAFU in human pluripotent-cell-derived neurons in vitro in fact drives splicing defects in DISC1 and ERRB4\(^{\text{51,55}}\). Moreover, these defects exactly quantitatively mimic those seen in schizophrenia patient cortex samples. This situation would mirror loss of function mutations in the GOMAFU locus.

**B.7.3 Alzheimer’s disease (AD)**

AD is a heritable neurodegenerative disorder that currently affects >1% of the living global population\(^{\text{34,87,19}}\). AD generally manifests in the elderly and is associated with progressively worsening dementia and memory loss, which eventually leads to the death of patients at an average of seven years post diagnosis\(^{\text{35,196}}\). A core molecular pathway driving progressive neurodegeneration in AD is thought to be the so-called ‘amyloid cascade’, in which toxic amyloid peptides accumulate and cause neuronal atrophy\(^{\text{202}}\).

The IncRNA BACE1-AS has been implicated in a positive feedback loop that drives progression of this amyloid cascade\(^{\text{76}}\). Briefly, BACE1-AS is transcribed antisense to and overlapping the BACE1 gene, which encodes a trans-membrane beta-secretase protein, whose dysregulation is well known to drive overproduction of pathogenic AB-42 peptides in AD. Functionally, BACE1-AS positively regulates BACE1 in vitro and in vivo, by binding to and stabilizing BACE1 mRNA via a 104-nucleotide region of perfect complementarity to exon 6 of the BACE1 mRNA. By inducing BACE1 expression, overexpression of BACE1-AS drives AB-
42 production in APP mutant HEK-SW cells. Interestingly, BACE1-AS expression is itself induced by elevated AB-42 peptide levels. Together these observations thus suggest a positive feedback loop, in which BACE1-AS drives overproduction of toxic AB-42 peptides, which then feed back to further induce BACE1-AS overexpression, accelerating amyloid accumulation. Consistent with this, BACE1-AS is expressed at 2-6 fold higher levels in AD patient brains relative to controls. Furthermore, perfusion of BACE1-AS targeting siRNAs into mice brains reduced both BACE1-AS and BACE1 expression. Reduction of BACE1 expression has independently been shown to ameliorate disease symptoms in animal models of AD, making this last result potentially therapeutically relevant. Nonetheless, it currently remains to be seen whether these insights translate into the human disease context.

The IncRNA BC1/200 has also been associated with AD disease progression. Although BC1/200 expression declines by >60% during normal aging, it was found to be significantly upregulated by up to 2.5-fold specifically in AD-affected brains, in regions including Brodmann area 9 and the hippocampus, relative to healthy age-matched controls. The magnitude of BC1/200 overexpression also correlated very strongly with the clinical dementia score in AD patient brains, and abnormal expression localization (non-somatodendritic) was observed in advanced AD brains. Although it is not yet clear whether these expression changes represent a cause or consequence of AD progression, these observations suggest another interesting candidate IncRNA for further exploration in AD research.

### B.7.4 Neuropathic pain

KCNA2-AS was introduced in the previous section on IncRNAs in neuronal plasticity. We refer the reader to that section for detailed mechanistic discussion of this IncRNA, but mention it again here for its role in neuropathic pain. Briefly, KCNA2-AS is induced in the dorsal root ganglion of rats in response to peripheral nerve injury. Specific induction of KCNA2-AS negatively regulates potassium channel subunit KCNA2 expression. Remarkably, forced overexpression of KCNA2-AS is sufficient to generate symptoms of neuropathic pain. Moreover, blocking KCNA2-AS expression with siRNAs attenuates the development of neuro-
pathic pain following peripheral nerve injury, such as spinal nerve ligation, or sciatic nerve axotomy. Thus, KCNA2-AS appears to be a key driver of neuropathic pain symptoms, and a potential therapeutic target to prevent human neuropathic pain.

B.7.5 Disease mechanisms outlook

All four disease-associated lncRNAs discussed in this section share a mechanism of action that hinges on close collaboration with specific protein partners. MSNPtAS regulates levels of MSN protein; GOMAFU regulates the activity of splicing proteins QKI, SRSF1 and SF3; BACE1-AS regulates beta-secretase expression; and KCNA2-AS regulates expression of the potassium channel KCNA2. In cases such as QKI and BACE1, mutation of these interacting proteins is also associated with the same disease state, potentially via disruption of the same pathway. However, the fact that GOMAFU and MSNPtAS map to independent GWAS risk loci suggests that they can play driving roles in disease progression. Awareness of lncRNA regulatory mechanisms may thus offer useful therapeutic targets – especially since in vivo manipulation of lncRNA expression is becoming possible. Proof of this concept will necessarily await clinical trials. Nonetheless, the outlook is optimistic, and continued attention to the many hundreds of other lncRNAs with either GWAS or transcriptional ties to neurological or psychiatric disorders will likely uncover additional interesting disease mechanisms.

B.8 lncRNAs in human brain evolution

B.8.1 Comparative genomics finds a lack of evidence for protein-driven human brain evolution

The genetic innovations responsible for phenotypic adaptation can be studied by comparing the genome sequences of related species. For example, the FOXP2 transcription factor, which is required for normal development of speech and language in humans, experienced a very strong selective sweep shortly after the
appearance of the Homo sapiens (rapid positive selection for specific amino acid substitutions relative to non-human primates, followed by strong purifying selection in human populations), suggesting a role in the evolution of human speech and language\textsuperscript{70}. ASPM and MCPH1 both also show signatures of positive selection in the human lineage, are expressed in dividing neural precursors, and cause microcephaly through loss of outer cortical layers when mutated, suggesting a role in human cortical expansion\textsuperscript{74,75}). AHI1 is involved in regulating axon pathfinding from the cortex to spinal cord, and also experienced positive selection in the human lineage, suggesting a role in regulating human-specific neural connectivity\textsuperscript{78}.

However, in contrast to these few specific examples of innovations in proteins that are associated with human-specific brain traits, the vast majority of nervous system proteins are in fact near perfectly conserved across diverse mammalian phyla\textsuperscript{249,164,81}. Moreover, systematic surveys of positive selection have uncovered a surprising lack of enrichment for accelerated amino-acid sequence changes (that imply advantageous variation) in protein-coding genes related to nervous system function in the human lineage relative to primates and rodents\textsuperscript{266,104}.

To reconcile this constancy with the obvious and dramatic changes in human brain anatomy and function relative to other mammalian species, it is common to appeal to elaboration of the combinatorial regulatory interactions, which control the spatiotemporal expression of nervous system genes during development, as a causal explanation of human brain adaptations\textsuperscript{168,223}. Indeed, accumulating evidence of significant changes in brain-region-specific gene expression across mammalian species implies that regulatory changes are widespread and likely to contribute to phenotypic novelty [reviewed in\textsuperscript{164}]. Another common explanation of this paradoxical result is that positive selection only needed to act on a small subset of nervous system protein coding genes, which were themselves sufficient to orchestrate human brain adaptations, while the majority of genes remained under intense negative selection due to their important and highly constrained functional roles\textsuperscript{109}.
B.8.2 Comparative genomics implicates lncRNAs in human brain evolution

Independent of protein-driven evolution, there is another attractive and complementary possibility: that human brain adaptations were driven by changes in non-protein-coding classes of genes, such as lncRNAs. In support of this, and in stark contrast to the highly-conserved repertoire of protein-coding-genes, recent comparative genomic analyses of mammalian lncRNAs have found that one third of human lncRNAs appear to be specific to the primate lineage\textsuperscript{64}, including hundreds of human specific lncRNAs\textsuperscript{281}. Many lncRNA loci have also experienced positive sequence selection during human evolution. To date, hundreds to thousands of lncRNA loci that are positively selected relative to other mammalian species\textsuperscript{164}, and 48 lncRNA loci that are positively selected within specific human populations\textsuperscript{91}, have been identified. Hundreds of these are independent of any protein-coding gene. An interesting example of a positively selected lncRNA is HARFi, which is expressed most-highly in Cajal-Retzius neurons during gestational weeks 7-19 of human neocortical development, a critical period of neuronal specification and migration, thus consistent with a new functional role in driving human-specific cortical development\textsuperscript{222}. Another interesting property of HARFi is that the positively selected regions of its locus are in fact highly conserved in other mammals, which might indicate that they indeed occur in a functional domain of the lncRNA to drive adaption. Overall the vast scale of this novel genetic information uniquely available to the developing human nervous system is difficult to ignore as a potential driver of human brain adaptations\textsuperscript{91,177,178,179}. It motivates us below to consider why we might see such a preponderance of genomic innovation in lncRNA genes relative to protein coding genes, and how this might relate to the emergence of human specific brain traits.

B.8.3 How novel lncRNAs could contribute to human-specific brain traits

(i) Increasing cortical size and cellular diversity. The human brain is approximately three times larger than chimpanzee brains, from which we diverged 7–8 million years ago, and about twice the size of pre-human hominids that lived approximately 2.5 million years ago\textsuperscript{44}. The most dramatically expanded human brain
region is the cortex, which serves as the seat of our higher cognitive functions. Relative to lower mammals, the human cortex contains higher overall cell numbers, a relative abundance of cortical interneurons, and some entirely new cell types, such as spindle cells, precocious predecessor cells, and fusiform cells. The human brain also has higher neuropil density, certain cortical regions that are not present in lower mammals, such as language centers, and a higher degree of asymmetry between hemispheres (reviewed in 64,128).

Thousands of new IncRNAs have appeared during primate nervous system evolution, where they are expressed in highly region-specific manner, consistent with a role in the spatiotemporal regulation of cellular identity. In fact IncRNAs serve as better markers for sub-populations of upper cortical neurons than protein coding genes197. They can moreover directly regulate cell fate choice in the developing brain through mechanisms analogous to those seen for EVF2, SIX3OS, and DLX1AS (see above), and can drive cellular identity changes when expressed ectopically94,169,231. LncRNA genes are therefore in principle ideal genetic substrates to have driven brain expansion and cellular diversification during human evolution. To gain direct evidence of this evolutionary function, researchers could in the future explore the effects of adding human-specific IncRNAs to ancestrally related mammalian brains, or look for human brain diseases associated with mutations in human-specific IncRNAs.

(ii) Enhancing learning through new regulation of neural plasticity The extent of synaptic interconnectivity, the organization of particular inter- and intra-regional circuit architectures, and the synaptic learning rules governing how specific circuits are formed, lost, strengthened, or weakened, are all known to influence the computational properties of the brain81. Human cognitive prowess may therefore be related to enhancement of these processes, though this idea awaits empirical support. LncRNAs directly modulate the duration and extent of synaptogenesis during development, and mature neural function, opening opportunities for new IncRNAs to influence human-specific circuit architectures. They are also integrated into activity-dependent regulatory circuits, where they dynamically control transcriptional, post-transcriptional, and translational changes, and could therefore refine human-specific regulation of neural plasticity. Primate specific IncRNAs such as BDNF-AS166, which is activity-dependent and regulates dendritic arborization (see
above), could be interesting subjects for future evolutionary studies.

(iii) **Scaffolding molecular interactions to improve inter-neuronal communication.** Control of signal-to-noise ratios is a major challenge associated with the expansion of highly interconnected neural networks that, if left unchecked, can offset any advantage in power of growing a larger brain\textsuperscript{145,166}. Certain molecular mechanisms have evolved that help combat these issues, such as the proteins HOMER and SHANK, which regulate interactions between signaling molecules within synapses, to improve the fidelity of neuronal signal transmission\textsuperscript{353}. The scaffold-like properties of IncRNAs make them ideal for analogous roles in modulating the biophysical properties of communicating neurons. One known example of a scaffold like IncRNA that acts at synapses, to influence inter-neuronal connectivity, is the BC1/BC200 IncRNA, which as noted above regulates stimulus-dependent translation of key plasticity-related genes and is required for normal brain function in mice\textsuperscript{45,323,327}. Other similar examples may be uncovered by future studies.

(iv) **Acting as new signaling molecules at synapses.** Another way to enhance the computational power of a neural network, of given size, is to allow for increasing numbers of independent messages with distinct meaning. Our brains are far more powerful for having multiple classes of neurotransmitters, such as glutamate, serotonin, dopamine, etc., than had they only utilized one such class of signal\textsuperscript{81}. Intercellular communication through vesicle-mediated transport of IncRNAs, small ncRNAs, and mRNAs, is increasingly being identified as an important physiological and developmental regulatory mechanism\textsuperscript{66,81,293}. Such mechanisms are known to operate between cells of the nervous system, but the functional consequences of this mechanism remain poorly understood. However, by transmitting IncRNAs across synapses, it would in principle be possible to tune local post-synaptic properties, for example by regulating post-synaptic translation of specific mRNAs, via IncRNAs such as BC1/200, or to regulate the global transcriptional state of the post-synaptic neuron, by IncRNAs that regulate transcription. If evolution has accessed this communication paradigm, IncRNAs may therefore have contributed to the evolution of cognition by enriching information encoded in neural communication.
B.8.4 Why might evolution have extensively employed lncRNAs as an adaptive genetic substrate?

Modularity is an organizational property that is characteristic of evolvable living systems\textsuperscript{90,141}. Modularity of functional RNA molecules has been theoretically demonstrated to spontaneously arise from environmental canalization under constant selective pressures\textsuperscript{7}. Indeed, metazoan lncRNAs are composed of modular functional domains, which each possess unique sets of DNA, RNA, or protein binding partners, that form regulatory linkages between pre-existing genetic and biochemical pathways (Box 1). TERC is a good example of such a lncRNA, and ranges from 200bp to 5kb depending on the set of modules spliced into the mature lncRNA from the multiexon gene locus\textsuperscript{165}.

The modular organization of lncRNAs is highly evolvable for at least three reasons. First, it allows each domain to explore new functional properties, such as new or altered binding partners, somewhat independently of the other functional domains. This is different to the case of proteins, where the overall structure and function of the protein macromolecule often depends more heavily on a large number of the constituent residues. Second, it allows parts of the lncRNA gene body that do not already contain functional domains to freely explore sequence space, and potentially evolve new domains. There may be some selective pressures on maintaining particular secondary structures, but this is dramatically lower than in protein coding genes, as evidenced by the much lower sequence conservation in lncRNA gene bodies relative to coding sequences. Finally, it allows new combinations of modules to be readily generated by transposition or recombination events\textsuperscript{125,136}.

Here we suggest that the evolvable nature of the emergent modular organization of lncRNA molecules helps to explain why lncRNA genes appeared and then rapidly expanded in modern metazoan species, and therefore furthermore supports a role for lncRNAs in driving the adaptive processes underlying human brain evolution.
B.9 Conclusion

The initial discovery of tens of thousands of lncRNAs that show exquisitely spatiotemporally specific expression patterns in the mammalian brain raises the obvious prospect that they are biologically meaningful, and begs the consequent question as to what their functional roles (if any) may be. Here we have reviewed emerging studies that have begun to address this question, and uncovered essential roles for lncRNAs in the development, plasticity, and disease of mammalian nervous systems. The basic and biomedical implications of these findings, especially with many thousands of other lncRNA transcripts that remain poorly understood, are abundantly clear. We have further synthesized these results within a comparative genomics framework, to argue that lncRNAs may furthermore represent a previously poorly appreciated driver of human brain adaptations responsible for the evolution of human cognition\textsuperscript{147,181,190}. We hope that these efforts encourage other researchers to share our enthusiasm and in future hold new awareness of the potential underlying roles of lncRNAs in their neural system of interest. Looking further forward, we anticipate that this momentum will extend to the emerging field of epitranscriptomics\textsuperscript{237}, as it seems clear that the molecular basis of the plasticity of the nervous system is intimately linked to the expansion of RNA editing and RNA modification, which have expanded during cognitive evolution, as well as DNA remodeling (by reverse transcriptase-linked DNA ‘repair’) and retrotransposon mobilization, which occur in the human brain\textsuperscript{11,41,180,181,237}. 
References


mental biology, volume 11 (pp. 133–186). Elsevier.


ferred from single-cell transcriptomics show a series of transitions through discrete cell states. *elife, 6*, e20487.


216


218


220


