THE INITIATION OF TRANSCRIPTION AND GASTRULATION IN THE *C. ELEGANS* EMBRYO

Sophia C. Tintori

A dissertation submitted to the faculty of the University of North Carolina at Chapel Hill in fulfillment of the requirements for the degree of Doctor of Philosophy in the Department of Biology in the College of Arts and Sciences.

Chapel Hill
2017

Approved by:

Robert P. Goldstein

Victoria L. Bautch

Frank L. Conlon

Robert J. Duronio

Corbin D. Jones

ABSTRACT

Sophia C. Tintori: The Initiation of Transcription and Gastrulation in the *C. elegans* Embryo
(Under the direction of Robert P. Goldstein)

During embryonic development, cells must establish their fates, change their composition and shape to perform metabolic or structural functions, and move around each other to form the shapes of the body. An outstanding challenge of developmental biology is to explain how cells enact these fundamental processes, and coordinate with each other to create a complex multicellular animal. A prevalent hypothesis for how this coordination is achieved is that each cell's fate and behavior is determined by a defined mixture of RNA transcripts. I created a resource documenting the full roster of transcripts present in each cell through the first few stages of *C. elegans* development. These embryonic cells are used as models to study a broad range of cell and developmental biology phenomena with a single-cell resolution. This new resource allows researchers to know, with a single-cell resolution, which transcripts are present and may be driving the diversity of biological phenomena taking place during these stages.

With this resource, I described cell-specific patterns in the initiation of zygotic transcription. I identified novel genes that are critical for development. Finally, I have identified a family of proteins that may be triggering a specific cell behavior— that of cell motility during gastrulation.

Beyond using this transcriptome dataset for my own research questions, I also developed an interactive data visualization tool. The goal of this tool is to ensure that the data are easily accessible to all interested researchers. Democratization of information is important both within

the scientific community as well as between the scientific community and the public. Toward that end, I produced several animations, short narrative videos, and a short documentary film to share scientific stories with scientists, students, grantors, and the general public.

ACKNOWLEDGEMENTS

I would like to thank my mentor, Bob Goldstein, for going out of his way to create a space where I could freely explore my ideas, both in the lab and in science communication projects. I have been incredibly lucky to have such an advocate as a mentor.

I'd like to thank my thesis committee: Vicki Bautch, Bob Duronio, Frank Conlin and Corbin Jones. Committee meetings were surprisingly some of my favorite memories of graduate school— thank you for making me feel like a valued and respected member of the scientific community, for your insightful feedback on my project, and for your support during complicated times.

To the Goldstein lab: Jessica Sullivan-Brown, Jenny Tenlen, Jenny Heppert, Chris Higgins, Dan Dickinson, Ari Pani, Thomas Boothby, Frank Smith, Mark Slabodnick and Tim Cupp. I thank you all for your comradery of the past six years. Your incredibly consistent passion, patience and humor have been my bedrock.

To the Lieb lab: Jason Lieb, Erin Osborne Nishimura, Tess Jeffers, Kohta Ikegami, Sebastian Pott, Colin Lickwar, Sheera Adar, Shezhad Sheikh, Laura Simmons Kovacs, Max Boeck, Jay Zhang and Asia Mieczkowska. This vibrant, exciting scientific family inspired me from the first day of my rotation. I cannot wait to see all the wonderful things you guys do.

To the Safe Spaces Committee: Kriti Sharma, Jenny Heppert, Alissa Brown, Laura Mudge, Katrina Kutchko and Laura Sligar. The work we did together was some of the work I'm

proudest of so far. I feel so lucky to have shared brain space with such a brilliant group of thoughtful, hardworking, compassionate women.

For being my tether and letting me lean on you during the last six years, I'd like to thank the incredible friends I've made while living in Chapel Hill: Rajeev Rajendran, Kat Stein, Jeremy Harris, Ethan Clauset, Brandon Joyce, Jonny Corndawg, Rachel Lewallen, Rosa Kneller, Patrick Golden, Jeff Rehnlund, Jon Page, Kim and Dan Marston and Mark Iosifescu. Thank you for making it fun.

And finally, for making me and trusting I would do something interesting with the opportunity, I'd like to thank my family: My mother, Mary Cybulski, my father, John Tintori, and my brother, Ray Tintori.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF ABBREVIATIONS

AB	Anterior cell of the 2-cell embryo
ABa	Anterior daughter cell of AB (4-cell stage)
ABal	Left daughter of ABa cell (8-cell stage)
ABalx	Two daughters of ABal (16-cell stage)
ABar	Right daugheter of ABa cell (8-cell stage)
ABarx	Two daughters of ABar (16-cell stage)
ABa-E-P_4	ABa, E, and P_4 descendants (100-cell stage)
ABp	Posterior daughter cell of AB (4-cell stage)
ABpl	Left daughter of ABp cell (8-cell stage)
ABplx	Two daughters of ABpl (16-cell stage)
ABpr	Right daughter of ABp cell (8-cell stage)
ABprx	Two daughters of ABpr (16-cell stage)
ABp+	16-32 AB descendants (100-cell stage)
ABp-	100-cell stage embryo minus the ABp descendants
ABxx	Four granddaughters of AB (8-cell stage)
ABxxx	Eight great-granddaughters of AB (16-cell stage)
BLAST	Basic local alignment search tool
C	Neuron, epidermis and muscle precursor, daughter of P_2 (8-cell stage)
Cx	Two daughters of C (16-cell stage)
C-in, Cxp, Cap&Cpp	4-8 posterior granddaughters of C, internalize during gastrulation (100-cell stage)

C-out, Cxa, Cpa&Caa	4-8 anterior granddaughters of C, not internalizing during gastrulation (100-cell stage)
C-	100-cell stage embryo minus the C descendants
D	Muscle precurson, daughter of P_3 (16-cell stage)
dsRNA	Double stranded RNA
D+	2-4 D descendants (100-cell stage)
D-	100-cell stage embryo minus the D descendants
E	Endoderm precursor cell, daughter of EMS (8-cell stage)
EMS	Endomesodermal precursor cell, daughter of P_1 (4-cell stage)
EMT	Epitherlial-Mesenchymal Transition
ERCC	External RNA Controls Consortium
EST	Expressed sequence tag
Ex	Two daughters of E (16-cell stage)
GFP	Green fluorescent protein
mNG	mNeonGreen
MS	Mesoderm precursor cell, daughter of EMS (8-cell stage)
MSx	Two daughters of MS (16-cell stage)
MS+	8-16 MS descendants (100-cell stage)
MS-	100-cell stage embryo minus the MS descendants
P_1	Posterior cell of the 2-cell embryo
P_2	Germ cell and mesoderm precursor, daughter of P_1 (4-cell stage)
P_3	Muscle and germ cell precursor, daughter of P_2 (8-cell stage)
P_4	Germ cell precursor, daughter of P_3 (16-cell stage)

PCA	Principle component analysis
RNAi	RNA interference
RPKM	Reads per kilobase of transcript per million mapped reads
scRNA-Seq	Single cell RNA sequencing
smFISH	Single molecule fluorescent *in situ* hybridization

CHAPTER 1: INTRODUCTION

1.1 Transcription During Early *C. elegans* Embryogenesis

During embryonic development, cells must establish their fates, move around each other to form structures such as organs, and change their composition and shape to perform metabolic or structural functions. An outstanding challenge of developmental biology is to explain how cells enact these fundamental processes and coordinate with each other to create a complex multicellular animal. A prevalent hypothesis for how this coordination is achieved is that each cell's fate and behavior is determined by a defined mixture of RNAs.

When mRNA transcription is inhibited in the *C. elegans* embryo via RNAi of *ama-1* (the large subunit of RNA polymerase II), the first visibly detectable defect is the failure of the gut precursors to gastrulate [Powell-Coffman et al. 1996]. Within the next two cell cycles, the mesodermal and germ line precursors also fail to gastrulate. The embryo then arrests before the 100 cell stage, consistent with the expectation that morphogenesis is not the only process in the early embryo that is controlled by transcription.

While the genetic triggers for gastrulation in the worm are unknown, the *ama-1* RNAi result suggests that it is transcriptionally regulated. This led my collaborators and me to hypothesize that it is possible to identify regulators of morphogenesis by investigating changes in the transcriptomes of gastrulating *C. elegans* cells at the time points leading up to gastrulation.

1.2 Genomic Approaches in the *C. elegans* Embryo

The *C. elegans* embryo is a powerful and well-established system for studying cell biology and development, and was chosen as a model organism in part because the entirety of development can be tracked with single-cell resolution (Sulston et al., 1983). The timing and orientation of every cell division, apoptotic event, and cell migration has been documented, and the exact lineal relationship of any cell to any other is known.

Many of the critical phenomena of early development are transcriptionally regulated in *C. elegans*, including morphogenesis and cell-fate specification (Edgar et al., 1994; Sommermann et al., 2010; Broitman-Maduro et al., 2006). Much of what we know about the genetics of these events has been gleaned from traditional genetic screens, but most genetic screens test only a single gene at a time, most often disrupting that gene in the whole embryo without spatial or temporal specificity. Therefore, this technique presents a blind spot for pleiotropic genes and genes with partially redundant functions (Wieschaus, 1997; Sawyer et al., 2011). Genomic approaches, such as high-throughput RNA sequencing, allow us to identify any gene whose transcript abundance correlates with morphogenesis, differentiation, or any other phenomena, regardless of challenges such as pleiotropy or redundancy.

Yet performing genomic studies in the *C. elegans* embryo with a high temporal resolution has been a challenge. Until recently, genomic protocols required collection of embryos in bulk, but *C. elegans* fertilization is staggered, rendering embryos asynchronous with one other. There is no practical system in place for culturing single cell types, therefore the only source of bulk biological material must be staged by hand or by fluorescence, which are both relatively imprecise, and are most often composed of mixed cell types (Baugh et al., 2003, Stoeckius et al., 2009).

The recent availability of reliable low-input RNA-seq protocols provides a solution to the challenge of purifying cell types from mixed-stage embryos (Svensson et al., 2017, Adiconis et al., 2013). A sample collected and sequenced from a single embryo can be precisely identified in time and space. These methods make it possible to generate comprehensive rosters of every transcript's abundance in an organism or tissue during key developmental events.

1.3 Current and Previous Work Documenting transcriptomes of the Early *C. elegans* Embryo

In this document, I report the dissection and sequencing of 40 distinct samples between the 1- and 100-cell stage embryo. During these stages, a broad range of developmental and cell biological phenomena take place, including cell fate specification, asymmetric cell division, the first gaps in the cell cycle, germ line specification, zygotic genome activation, and the beginning of morphogenesis. From these RNA-seq libraries, my collaborators and I have described transcriptomes for 30 distinct cell types. Because each transcriptome was generated from samples taken from a single embryo, the samples can be precisely defined in time and space. We used these transcriptomes to address previously unanswered questions about the differential activation of the zygotic genome in each cell, to describe spatially dynamic gene expression, to identify previously unknown genes that are critical for development, and finally to identify candidate genes for the regulation of morphogenesis.

Understanding the full suite of mRNAs expressed in the *C. elegans* embryo has long been of interest. Whole-embryo mRNA time courses revealed that thousands of genes are dynamically regulated at these early stages (Baugh et al., 2003, 2005). Aided by advances in low-input RNA-seq technology of the last few years, researchers have interrogated the transcriptomes of the embryo by manually dissecting cells and performing RNA-seq. Due to the difficulty of

identifying cells once they are dissected, only the 2-cell stage embryo has been sequenced at an entirely single-cell resolution (Hashimshony et al., 2012, 2015; Osborne Nishimura et al., 2015). Researchers in one study performed transcript profiling of some single cells and some clusters of cells from later stages (Hashimshony et al., 2015). In my PhD research, I contributed to this pursuit by performing hand-dissections of complete sets of single cells from individual embryos and sequencing their mRNA. For samples from early stages, where fluorescent markers were not available to identify each cell collected, I developed a unique computational strategy to assign cell identifies to each transcriptome *post-hoc*. For later stages, where cells were small and single-copy fluorescent markers were dim, I performed high-precision manual dissections on the stage of a spinning disc confocal microscope.

1.4 An Introduction to Morphogenesis

During early embryogenesis, cells move around each other to form the shapes of the body— a process called morphogenesis. Failure of morphogenesis can result in birth defects (such as neural tube defects and congenital heart disease)(Wallingford et al., 2013, Srivastava 2006). Conversely, the inappropriate activation of this process can contribute to the spread of cancer (as is known from studies of snail in EMT and tumor progression)(Blanco et al., 2002, Batlle et al., 2000).

During phenomena such as neurulation, gastrulation, lung branching and heart looping, entire swaths of cells simultaneously move or change their shape, thereby reshaping tissues that can be orders of magnitude larger than any of the individual cells. Though each of these events takes place in different cell types, they use many of the same physical mechanisms. For example, both neurulation in the frog and gastrulation in the worm rely on apical constriction.

Furthermore, some of the genes implicated in neural tube closure defects in mammals lead to gastrulation defects when disrupted in *C. elegans* (Sullivan-Brown et al., 2016). These morphogenetic events involve physical changes in the cell, such as assuming polarity, modulating adhesion, and locally activating the cytoskeleton (Purnell 2013).

It has been a long-standing challenge to identify the links between genetic patterning of the embryo and the mechanics of morphogenesis (Wieschaus, 1997, Sawyer et al., 2011, Gilmour et al., 2017). A small group of genes whose embryonic transcription trigger a morphogenetic behavior are known, setting a precedent for the type of genes we might expect in *C. elegans*. These known genes include snail, which has been shown to be responsible for epithelial-mesenchymal transition (EMT) in many animals including flies (Leptin, 1991) mice (Carver et al., 2001), chick (del Barrio and Nieto, 2002) and urchin (Shu-Yu and McClay, 2007, Nieto, 2002), and shroom3, an actin-binding protein that recruits ROCK to apical junctions (Hildebrand and Soriano, 1999, Nishimura and Tackeichi, 2008) and whose expression is sufficient to induce apical constriction in polarized epithelial cells (Haigo et al., 2003). We hypothesized that more genes like this— master regulatory genes for diverse morphogenetic events within an embryo and even across species— exist and could be uncovered with a strategic RNA-seq approach.

1.5 An Introduction to Gastrulation in *C. elegans*

The first described morphogenetic event in the *C. elegans* embryo is the internalization of the E lineage (intestine precursor cells) at the 26-cell stage, marking the beginning of gastrulation (Nance et al, 2005). After the E cells internalize, several other cell types follow, including muscle precursors and the germ line. My collaborators and I were interested in

studying the transcriptomes of internalizing cells, but studying the transcriptome of only the E lineage would lead to ambiguous results, as transcripts that regulate morphogenesis would be indistinguishable from transcripts that regulate endoderm fate. Rather than studying just one internalizing cell type, we decided to compare transcriptomes across several internalizing cell lineages as they adopt different fates but initiate similar morphogenetic processes. We hypothesized that we would find transcripts that become enriched over time in cell types that are preparing to internalize, but not in cell types that do not internalize. Furthermore, we hypothesized that those transcripts would be significant for the regulation of this morphogenetic event.

REFERENCES

Adiconis, X., Borges-Rivera, D., Satija, R., DeLuca, D.S., Busby, M.A., Berlin, A.M., Sivachenko, A., Thompson, D.A., Wysoker, A., Fennell, T., Gnirke, A., Pochet, N., Regev, A., & Levin, J.Z. (2013) Comparative analysis of RNA sequencing methods for degraded or low-input samples. Nature Methods 10(7):623–629.

Ahn D.G., Kourakis M.J., Rohde L.A., Silver L.M. & Ho R.K. (2002). T-box gene tbx5 is essential for formation of the pectoral limb bud. Nature 417(6890):754-8.

Andachi, Y. (2004). Caenorhabditis elegans T-box genes tbx-9 and tbx-8 are required for formation of hypodermis and body-wall muscle in embryogenesis. Genes Cells 9(4):331-44.

Batlle, E., Sancho, E., Francí, C., Domínguez, D., Monfar, M., Baulida, J. & García De Herreros, A. (2000). The transcription factor snail is a repressor of E-cadherin gene expression in epithelial tumour cells. Nat Cell Biol 2(2):84-9.

Baugh, L. R., Hill, A. A., Claggett, J. M., Hill-Harfe, K., Wen, J. C., Slonim, D. K., Brown, E.L. and Hunter, C.P. (2005). The homeodomain protein PAL-1 specifies a lineage-specific regulatory network in the C. elegans embryo. Development, 132(8), 1843–1854. http://doi.org/10.1242/dev.01782

Baugh, L. R., Hill, A. A., Slonim, D. K., Brown, E. L., & Hunter, C. P. (2003). Composition and dynamics of the Caenorhabditis elegans early embryonic transcriptome. Development, 130(5), 889–900.

Blanco, M.J., Moreno-Bueno, G., Sarrio, D., Locascio, A., Cano, A., Palacios, J. & Nieto, M.A. (2002). Correlation of Snail expression with histological grade and lymph node status in breast carcinomas. Oncogene 21(20):3241-3246. http://doi.org/10.1038/sj.onc.1205416

Broitman-Maduro, G., Lin, K. T.-H., Hung, W. W. K., & Maduro, M. F. (2006). Specification of the C. elegans MS blastomere by the T-box factor TBX-35. Development, 133(16), 3097–3106. http://doi.org/10.1242/dev.02475

Carver E.A., Jiang R., Lan Y., Oram K.F. & Gridley T. (2001) The mouse snail gene encodes a key regulator of the epithelial-mesenchymal transition. Mol Cell Biol. 21(23):8184-8.

Cebra-Thomas J.A., Bromer J., Gardner R., Lam G.K., Sheipe H., Gilbert S.F. (2003). T-box gene products are required for mesenchymal induction of epithelial branching in the embryonic mouse lung. Developmental Dynamics 226(1) 82–90.

del Barrio, M.G. and Nieto, M.A. (2002). Overexpression of Snail family members highlights their ability to promote chick neural crest formation. Development 129(7):1583-93.

Edgar, L. G., Wolf, N., & Wood, W. B. (1994). Early transcription in Caenorhabditis elegans embryos. Development, 120(2), 443–451.

Garrity, D.M., Childs, S. & Fishman, M.C. (2002). The heartstrings mutation in zebrafish causes heart/fin Tbx5 deficiency syndrome. Development 129(19):4635-45.

Gilmour, D., Rembold, M. & Leptin, M. (2017). From morphogen to morphogenesis and back. Nature. 541(7637):311-320. http://doi.org/10.1038/nature21348

Haigo, S.L., Hildebrand, J.D., Harland, R.M. & Wallingford, J.B. (2003). Shroom induces apical constriction and is required for hingepoint formation during neural tube closure. Curr Biol., 13(24), 2125-37.

Hildebrand, J.D. and Soriano, P. (1999). Shroom, a PDZ domain-containing actin-binding protein, is required for neural tube morphogenesis in mice. Cell 99: 485-497. http://doi.org/10.1016/S0092-8674(00)81537-8.

Hashimshony, T., Feder, M., Levin, M., Hall, B. K., & Yanai, I. (2015). Spatiotemporal transcriptomics reveals the evolutionary history of the endoderm germ layer. Nature, 519(7542), 219–222. http://doi.org/10.1038/nature13996

Hashimshony, T., Wagner, F., Sher, N., & Yanai, I. (2012). CEL-Seq: single-cell RNA-Seq by multiplexed linear amplification. Cell Reports, 2(3), 666–673. http://doi.org/10.1016/j.celrep.2012.08.003

Leptin, M. (1991) twist and snail as positive and negative regulators during Drosophila mesoderm development. Genes Dev 5(9):1568-76.

Nance, J., Lee, J.-Y., & Goldstein, B. (2005). Gastrulation in C. elegans. WormBook : the Online Review of C. Elegans Biology, 1–13. http://doi.org/10.1895/wormbook.1.23.1

Nieto MA. (2002). The snail superfamily of zinc-finger transcription factors. Nat Rev Mol Cell Biol 3:155-66.

Nishimura, T. and Takeichi, M. (2008). Shroom3-mediated recruitment of Rho kinases to the apical cell junctions regulates epithelial and neuroepithelial planar remodeling. Development 135: 1493-1502. http://doi.org/10.1242/dev.019646.

Osborne Nishimura, E., Zhang, J. C., Werts, A. D., Goldstein, B., & Lieb, J. D. (2015). Asymmetric transcript discovery by RNA-seq in C. elegans blastomeres identifies neg-1, a gene important for anterior morphogenesis. PLoS Genetics, 11(4), e1005117. http://doi.org/10.1371/journal.pgen.1005117

Powell-Coffman, J.A., Knight, J. & Wood, W.B. (1996). Onset of C. elegans gastrulation is blocked by inhibition of embryonic transcription with an RNA polymerase antisense RNA. Dev Biol., 178(2):472-83.

Purnell B.A. (2013). Getting into Shape; Introduction to Special Issue. Science 340(6137) 1183. http://doi.org/10.1126/science.340.6137.1183

Sawyer, J. M., Glass, S., Li, T., Shemer, G., White, N. D., Starostina, N. G., Kipreos, E.T., Jones, C.D. & Goldstein, B.(2011). Overcoming redundancy: an RNAi enhancer screen for

morphogenesis genes in Caenorhabditis elegans. Genetics, 188(3), 549–564. http://doi.org/10.1534/genetics.111.129486

Showell C., Binder O., Conlon F.L. (2004). T-box genes in early embryogenesis. Dev Dyn. 229(1):201-18.

Shu-Yu W. and McClay D.R. (2007) The Snail repressor is required for PMC ingression in the sea urchin embryo. Development 134, 1061-1070. http://doi.org/10.1242/dev.02805

Sommermann, E. M., Strohmaier, K. R., Maduro, M. F. & Rothman, J. H. (2010). Endoderm development in Caenorhabditis elegans: the synergistic action of ELT-2 and -7 mediates the specification → differentiation transition. Developmental Biology, 347(1), 154–166. http://doi.org/10.1016/j.ydbio.2010.08.020

Srivastava D. (2006). Making or Breaking the Heart: From Lineage Determination to Morphogenesis. Cell 126(6) 1037–1048. http://doi.org/10.1016/j.cell.2006.09.003

Stoeckius, M., Maaskola, J., Colombo, T., Rahn, H.-P., Friedländer, M.R., Li, N., Chen, W., Piano, F. & Rajewsky, N. Large scale sorting of C. elegans embryos reveals the dynamics of small RNA expression. Nat Methods. 2009 Oct; 6(10): 745–751.

Sullivan-Brown, J.L., Tandon, P., Bird, K.E., Dickinson, D.J., Tintori, S.C., Heppert, J.K., Meserve, J.H., Trogden, K.P., Orlowski, S.K., Conlon, F.L. & Goldstein, B. (2016). Identifying Regulators of Morphogenesis Common to Vertebrate Neural Tube Closure and Caenorhabditis elegans Gastrulation. Genetics 202(1):123-39. http://doi.org/10.1534/genetics.115.183137.

Sulston, J. E., Schierenberg, E., White, J. G., & Thomson, J. N. (1983). The embryonic cell lineage of the nematode Caenorhabditis elegans. Developmental Biology, 100(1), 64–119.

Svensson, V., Vento-Tormo, R. and Teichmann, S.A. (2017). Exponential scaling of single-cell RNA-seq in the last decade. BioArXiv pre-print. arXiv:1704.01379v2.

Wallingford J.B., Niswander L.A., Shaw G.M. & Finnell R.H. (2013). The Continuing Challenge of Understanding, Preventing, and Treating Neural Tube Defects. Science 339 (6123), 1222002. http://doi.org/10.1126/science.1222002

Wieschaus, E. (1997). Nobel Lecture. In N. Ringertz (Ed.), Nobel Lectures. Singapore.

Wilson V. and Conlon F.L. (2002). The T-box family. Genome Biol. 2002;3(6):REVIEWS3008.

CHAPTER 2: A TRANSCRIPTIONAL LINEAGE OF EARLY *C. ELEGANS* EMBRYOGENESIS

The work described in this chapter was published in 2016.

A transcriptional lineage of early C. elegans development.
Tintori, SC, Osborne Nishimura, E, Golden, PT, Lieb, JD, Goldstein, B.
Developmental Cell. 2016 Aug 22;38(4):430-444. doi: 10.1016/j.devcel.2016.07.025

The central idea and experiments described in this paper were conceived by me, with guidance from Jason Lieb and Bob Goldstein. All single-molecule fluorescent in situ hybridization experiments (in Figures 2.3 and 2.5) were performed by Erin Osborne Nishimura, and all coding for the interactive data visualization program (Figure 2.7) was performed by Patrick Golden. The manuscript was written by me, with feedback from all collaborators. Supplementary figures and tables are not included in this thesis, but can be found through the Developmental Cell website.

During embryonic development, cells must establish fates, morphologies and behaviors in coordination with one another to form a functional body. A prevalent hypothesis for how this coordination is achieved is that each cell's fate and behavior is determined by a defined mixture of RNAs. Only recently has it become possible to measure the full suite of transcripts in a single cell. Here we quantify the abundance of every mRNA transcript in each cell of the *C. elegans* embryo up to the 16-cell stage. We describe spatially dynamic expression, quantify cell-specific differential activation of the zygotic genome, and identify genes that were previously unappreciated as being critical for development. We present an interactive data visualization tool that allows broad access to our dataset. This genome-wide single-cell map of mRNA abundance, alongside the well-studied life history and fate of each cell, describes at a cellular resolution the mRNA landscape that guides development.

2.1 INTRODUCTION

An outstanding challenge of developmental biology is to explain how differential gene expression promotes the fundamental processes of embryonic development. Such processes include determining the fate of each cell, moving cells relative to each other to produce structures such as organs, and changing the composition and shape of each cell to perform metabolic or structural functions. Genomic approaches developed over the past decade have made it possible to generate comprehensive rosters of every transcript's abundance in an organism or tissue during key developmental events. In this study, we have measured the abundances of all mRNAs in each cell of the early *C. elegans* embryo. In doing so, we have quantified the divergence of the genetic expression of these cells as they begin to perform diverse functions in the embryo.

The *C. elegans* embryo is a powerful and well-established system for studying cell biology and development (Figure 2.1A), and was chosen as a model organism in part because the entirety of development can be tracked with single-cell resolution (Sulston et al. 1983). The timing and orientation of every cell division, apoptotic event, and cell migration has been documented, and the exact lineal relationship of any cell to any other is known. Yet performing genomic studies with a matching resolution has been a challenge. Until recently, genomic protocols required collection of embryos in bulk, but *C. elegans* fertilization is staggered, rendering embryos asynchronous with each other. There is no practical system in place for culturing single cell types, leaving the only source of bulk biological material imprecisely staged samples that are usually composed of mixed cell types. Low-input RNA-sequencing (RNA-seq) methods developed within the last five years offer a solution to the genomics problem; a single *C. elegans* cell can be precisely identified and defined both in space and time.

Understanding the full suite of mRNAs expressed in the *C. elegans* embryo has long been of interest. Whole-embryo mRNA timecourses revealed that thousands of genes are dynamically regulated at these early stages (Baugh et al. 2003; Baugh et al. 2005). Aided by advances in low-input RNA-seq technology of the last few years, researchers have interrogated the transcriptomes of the embryo by manually dissecting cells and performing RNA-seq. Due to the difficulty of identifying cells once they are dissected, only the 2-cell stage embryo has been sequenced at an entirely single-cell resolution (Hashimshony et al. 2012; Hashimshony et al. 2015; Osborne Nishimura et al. 2015). One study has performed transcript profiling of some single cells and some clusters of cells from later stages (Hashimshony et al. 2015). In this study we have sequenced each cell of an individual embryo in replicate for embryos up to the 16-cell stage. We hand-dissected complete sets of single cells from each embryo, and developed a unique strategy for identifying the dissected cells.

Many of the interesting phenomena of early development are transcriptionally regulated in *C. elegans*, including morphogenesis and cell fate specification (Edgar et al. 1994; Sommermann et al. 2010; Broitman-Maduro et al. 2006). Much of what we know about the genetics of these events has been gleaned from traditional genetic screens, which have a blind spot for pleiotropic genes and genes with partially redundant functions (Wieschaus 1997; Sawyer et al. 2011). With high-throughput sequencing, we can identify the genes whose transcript abundances correlate with morphogenesis, differentiation, or other phenomena, regardless of a gene's possible pleiotropy or redundancy.

Here we present a transcriptional lineage of early *C. elegans* development – a map of all transcripts in each cell through the first stages of development. We generated this map by performing single cell RNA-seq (scRNA-seq) on each cell from the zygote to the 16-cell stage.

We address previously unanswered questions about the differential activation of the zygotic genome in each cell, describe spatially dynamic gene expression, and identify previously unknown and potentially redundant genes that are critical for development. Finally, we introduce a publicly available interactive data visualization tool that we developed to maximize the usefulness of our dataset to the scientific community.

2.2 RESULTS

Transcriptome Diversity Among Cells of the Embryo Increases Over Time

Each cell at each stage in the early *C. elegans* embryo has a name, a known life history and fate, and is identifiable by its position relative to other cells (Sulston et al. 1983). We performed scRNA-seq on manually-dissected, individual cells from 1-, 2-, 4-, 8- and 16-cell stage embryos, with a minimum of 5 replicates for each sample (Figure 2.1B). We note that due to asynchronous cell divisions there is no true 16-cell stage, but we use this term for convenience (details in Experimental Methods). We sequenced the mRNA of each cell separately, knowing which embryo the cell came from but not knowing its identity. We then used its transcript profile to identify its cell type *post hoc*. Cell size and cell division timing gave us some clues of the identities of 19 of the 31 cell types. For example, all the anterior (AB descendant) cells at the 8- and 16-cell stages divide in synchrony with each other, and the germ cell precursors at the 2- to 8-cell stages are considerably smaller than the rest of the cells (purple in Figure S1, and Extended Methods). These visual clues provided independent support for the results of our *post hoc* cell identity assignments (Figure S1 and described below).

In total, we generated 219 transcriptomes, describing quantitative expression levels for 8,575 detected genes (>25 RPKM). We aggregated data from cells of the same embryo to

calculate whole-embryo statistics. To calculate the mass of mRNA in each cell, and thereby each whole embryo, we used spike-in controls from the External RNA Control Consortium (Baker et al. 2005). The mass of mRNA detected was relatively constant between stages, though embryos of later time points showed higher variability (Figure 2.1C). Among 31 whole embryos, five embryos had an mRNA mass more than one standard deviation above or below the average and were excluded from further analysis (details in Table S1). To evaluate changes in transcriptome complexity over time in individual cells and in whole-embryos, we calculated the number of mRNA species detected in each single-cell transcriptome and whole-embryo aggregation (Figure 2.1D,E). We noticed an increase in transcriptome complexity in whole embryos over time (>25 RPKM in any contributing cell), but a decrease in complexity in individual cells. The increase in whole-embryo complexity could be due to either cell-specific activation of the zygotic genome, or to the fact that a larger number of single-cell libraries constitute the whole embryo total at later stages, potentially allowing for fewer false negatives when compared to the small number of transcriptomes that make up whole-embryo values at earlier stages.

Before we could test the validity of the transcriptomes generated, we first needed to identify the cell type of origin for each transcriptome.

Posterior Cells of the Embryo Have Distinct Signatures Involving Hundreds of Cell-Specific Transcripts

Many of the cell types we sampled are enriched for transcripts of one or a few known marker genes, which we were able to use to assign identities to our transcriptomes. A multi-gene clustering approach has been shown to be more effective at grouping replicates of a cell type than a single- or few-gene approach (Björklund et al. 2016; Jaitin et al. 2014; Grün et al. 2015; Satija et al. 2015). We used an iterative Principle Component Analysis (PCA) strategy (described

below) to group transcriptomes by cell type, thereby collapsing our 219 transcriptomes down to 18 groups of identical or related cell types. We then used known marker genes to assign identities to each of these 18 groups (Figure 2.2). The 18 groups were P_0, AB, P_1, ABa, ABp, EMS, P_2, ABxx (granddaughters of AB), MS, E, C, P_3, ABxxx (great granddaughters of AB), MSx (daughters of MS), Ex (daughters of E), Cx (daughters of C), D and P_4. Some of the groups that contained multiple cell types were later sorted into more specific groups (Figure 2.3).

To filter for informative genes to use in our PCAs, we designed an algorithm to select genes that are reproducibly differentially enriched between cells of the embryo (details in Experimental Procedures). To group replicates of each cell type together, we performed a PCA on all transcriptomes of a given stage using just those filtered genes. We inspected plots of the first and second principle components for distinct groups consisting of one transcriptome from each embryo, which suggest grouping by shared cell-specific features (Figure 2B,E,I,N). We interpreted a group with exactly one cell from each embryo as comprising the replicates of a single (albeit unknown) cell type.

Each PCA tended to isolate only the most dramatically distinct cell types (Figure 2.2E,I,N). To then identify cell types with more subtle distinguishing features, we removed the transcriptomes that had already clustered out into independent groups, and re-ran the gene selection algorithm and PCA with just the remaining cells. In this way, we continuously enhanced our resolution and split groups of cells off based on increasingly subtle differences (Figure 2.2F,J,K,O,P; arrows show the cluster of remaining transcriptomes that were put through the next PCA iteration). We chose this iterative PCA approach because it allowed us to take advantage of a unique feature of the *C. elegans* embryo: Each embryo sampled from a given stage generated an identical number of transcriptomes, representing exactly the same set of cell

types. Many transcriptome clustering methods define clusters of unspecified size (Yan et al. 2013; Jaitin et al. 2014; Grün et al. 2015; Zeisel et al. 2015), but for this experiment it was most informative to identify groups consisting of exactly one transcriptome from each embryo (see Discussion). The simplest way to achieve this was to inspect the results of a PCA plot for isolated groups of transcriptomes that consisted of one transcriptome from each replicate (Figure 2.2).

The cells of the 2-cell stage embryo (AB and P_1) have noticeably different sizes, which allowed us to identify these cells during sample collection. We were able to use this previous knowledge to test the accuracy of our gene selection algorithm and PCA approach. We found that our strategy did in fact allow us to independently and accurately distinguish between these two cell types; all AB cells fell on one side of the first principle component, while all P_1 cells fell on the other side of the principle component (Figure 2.2B). The germ cell precursors in subsequent stages (P_2 at the 4-cell stage and P_3 at the 8-cell stage) were noticeably smaller than the others and so were also identified upon collection. These cells successfully segregated from the other cells types by our algorithm and PCA (Figure 2.2E,I). The independent identification of these cells as replicates of each other further validated our algorithm as an effective unsupervised method for selecting informative genes.

To assign identities to groups of cells distinguished by PCA, we examined genes that are known to be expressed in specific cell types. For example, *med-2* is known to be expressed in EMS at the four-cell stage (Maduro et al. 2001). Our transcriptome data shows high *med-2* levels exclusively and robustly in one distinct group of replicates at the four-cell stage (Figure 2.2G). Based on these observations, we concluded that this cluster consists of the EMS transcriptomes. Similarly, using known markers of cell identity, we verified AB and P_1 cells at the 2-cell stage,

AB daughter cells (ABa and ABp, referred to collectively here as ABx) and P_2 cells at the 4-cell stage, MS, E, C, P_3 and AB granddaughter cells (ABxx) at the 8-cell stage, and MS daughters (MSx), E daughters (Ex), C daughters (Cx), D, P_4, and AB great-granddaughters (ABxxx) at the 16-cell stage (Figure 2.2, Extended Methods).

Anterior Cells of the Embryo Were Indistinguishable from Each Other by an Unsupervised Multi-Gene Approach, but Show Differential Enrichment of Notch Target Gene mRNAs

For both the 8-cell and 16-cell stages, our PCA approach did not visibly distinguish the descendants of AB from each other (Figure 2.2K). These results indicate that the transcriptomes of AB descendants at these stages were very similar to each other. This is consistent with the fact that very few genes are known to be differentially expressed between these cells (Priess 2005). To distinguish between these transcriptomes, we examined them for transcripts of a few genes whose proteins are known to be differentially expressed between these cells, namely members of the notch signaling pathway, *hlh-27*, *ref-1* and *tbx-38* (Neves & Priess 2005). We queried all transcriptomes of the AB descendants at the 8- or 16-cell stages for transcripts of these three genes, and found that they offered enough information to partition these transcriptomes into four cell types at the 8-cell stage and four pairs at the 16-cell stage (Figure 2.3B,F, Extended Methods).

To match each hand-sorted group of transcriptomes to a specific cell identity, we performed single molecule FISH (smFISH) on these notch targets in intact 8- and 16-cell embryos. We analyzed micrographs to determine which cell of the embryo expressed each of the distinct combinations of notch targets seen in our data. At the 8-cell stage *hlh-27* transcripts were the most highly enriched in the ABpl and ABpr cells, *ref-1* transcripts were enriched in ABpl cells, and *tbx-38* transcripts were detected at very low levels primarily in ABal (Figure 2.3C,D).

At the 16-cell stage *hlh-27* was enriched in all AB descendants except the ABalx (ABala and ABalp) cells, *ref-1* was detected in ABarx and ABprx cells, and *tbx-38* was detected in ABalx and ABarx cells (Figure 2.3G,H). This smFISH data in combination with the scRNA-seq data for these notch targets allowed us to sort and identify transcriptomes into the four cell types at the 8-cell stage (ABal, ABar, ABpl and ABpr), and into four pairs of cell types at the 16-cell stage (ABalx, ABarx, ABplx and ABprx; Extended Methods).

The notch targets mentioned above are critical for cell fate specification of these anterior cells, and are activated via signaling from neighboring cells (Priess 2005). A previous study that sequenced all AB descendants together after allowing them to grow outside of their native embryonic environment showed no *hlh-27* expression in these cells, suggesting that key fate-determining signaling events may have been prevented (Hashimshony et al. 2015). This indicates that processing cells around 10 minutes after dissection, as we did, produces results that more accurately reflect the biology of intact embryos.

Together, our data reveal that the transcriptomes of AB descendants are almost indistinguishable from one another except for transcripts of a few genes, whereas P_1 descendants show hundreds of differences from one another. Some pairs of cells were ultimately indistinguishable from each other by our method, but because each cell was sequenced independently, this decreased resolution reflects the biology of the cells.

The Transcriptional Lineage Expands Upon Known Gene Expression Patterns During Development and Increases Their Resolution

Having assigned cell identities to each transcriptome in our dataset, we first confirmed that the data and our identity assignments reflected certain known expression patterns. We queried our dataset for expression patterns of *sdz-38* (which encodes a putative zinc finger

protein that is expressed in the MS cell; Robertson et al. 2004), *tbx-37* (a T-box transcription factor found in ABa descendants; Neves & Priess 2005), *ceh-51* (a homeodomain protein expressed in the MS lineage; Broitman-Maduro et al. 2009), *elt-7* (a GATA-type transcription factor that induces gut specification in the E descendants; Sommermann et al. 2010), *cwn-1* (a wnt ligand expressed in the C and D cells; Gleason et al. 2006), and *cey-2* (a putative RNA binding factor restricted to the germ line; Seydoux & Fire 1994). None of these genes were used to previously identify each cell type (Figures 2.2 and 2.3), so were able to use their expression patterns to independently test the validity of our data and our cell assignments. Our scRNA-seq data reflect the expected patterns for all six of these genes (Figure 2.4A, key in Figure 2.1F), and additionally quantify their expression in each cell, as well as that of the other 8,569 detected genes (Figure 2.4B).

Low-input transcriptomes for some of these cells, including AB and P_1, have previously been generated (Hashimshony et al. 2012; Osborne Nishimura et al. 2015; Hashimshony et al. 2015), as have whole-embryo microarray timecourses of *C. elegans* development (Baugh et al. 2003; Baugh et al. 2005). We compared our scRNA-seq data to data from two previous studies (Hashimshony et al. 2015, Osborne Nishimura et al. 2015) that each used different methods to sequence mRNA from AB and P_1 cells (2-cell stage). We calculated enrichment index values for each gene (a product of the gene's AB/P_1 fold change and the gene's average expression, Experimental Procedures). To measure the agreement between each study, we compared the enrichment index values calculated from each study's data. All studies were positively correlated with one another to similar extents, and the correlation increased when only significantly differentially enriched genes were compared (Figure S3A).

We analyzed data from the 2005 whole-embryo microarray timecourse (Baugh et al. 2005) to test if our low-input transcriptomes reflect the patterns identified by a higher-input but lower-sensitivity experiment. We searched for genes whose transcript levels either increased or decreased by twofold over time in the microarray data and identified 1,935 and 2,164 genes respectively. Transcripts whose levels increased or decreased in our dataset included 91% and 97% of those detected in the earlier dataset. In addition, we identified 7,763 other transcripts that increased or decreased over time, many of which had very low expression levels, presumably undetectable by microarray, and 1,053 for which there was no microarray probe in the previous experiment (Figure S3B). This result suggests that even though the transcriptomes we present here were generated from just picograms of mRNA, they capture the patterns described by a higher-input method, but with much greater sensitivity and resolution.

Transcriptional Dichotomy Between Germ Cells and the Soma

Visualization of transcript levels for all 8,575 genes detected across all cell types revealed three distinct trends of gene expression (Fig. 4B): First were transcripts only detected in subsets of cell types, suggesting cell-specific transcription (Fig 4B top). Second were transcripts present in the zygote that then became detected at lower levels over time in an embryo-wide fashion, suggesting global mRNA degradation (Fig. 4B center). Third were transcripts that were differentially abundant between somatic cells and germ cell precursors (Fig 4B bottom). Within this third group, in some cases transcripts became undetectable over time in the somatic cells but remained detectable in the germ cell and the immediate sister of the germ cell (as in *daz-1*, a gene required for oogenesis; Karashima et al. 2000, Figure 2.4C). In other cases, genes became detectable over time in the somatic cells, while remaining undetectable in germ cells and their

sisters (as in *skr-10*, a core component of the ubiquitin-ligase complex; Yamanaka et al. 2002; Figure 2.4D). Transcriptional quiescence in the germline is a feature that many organisms share (Deshpande et al. 2004; Cheung & Rando 2013). Our dataset quantitatively illustrates this phenomenon genome-wide.

Differential Activation of the Zygotic Genome Among Cell Lineages

Fundamental events of embryonic development start earlier in the *C. elegans* embryo than in many other model organisms. Cell-fate determining steps begin as early as the 2-cell stage, and gastrulation begins at the 26-cell stage. Within the embryo, certain cells engage in these events earlier than others. For example, gastrulation begins earliest in the E descendants and follows later in other cells (Nance et al. 2005). By further example, at the 16-cell stage the P_1 descendants (which we will refer to as posterior cells) include 4 cells that are already restricted to a single fate, while none of the AB descendants (which we will refer to as anterior cells) are as fully fate-restricted (Figure 2.1A). Based on this, we hypothesized that transcriptomes change more dramatically in the more fate-restricted posterior cells than the anterior cells. To quantify the extent to which transcriptomes of each lineage change over time, we asked how many genes were detected as having increased or decreased transcript levels in each cell when compared to the cell's parent. We found a higher number of both increasing and decreasing transcript levels in the non-germ descendants of the P_1 cell (Figure 2.4E,F) than in AB descendants, supporting our hypothesis that there is more dynamic gene regulation in these cells than in the AB descendants. We wondered whether this apparent increased dynamism of gene regulation (number of transcripts increasing or decreasing in abundance) in the non-germ posterior cells could be related to other features of these cells, such as greater mass of mRNA (Figure 2.4I), greater

transcriptome complexity (Figure 2.4J), or longer cell cycle (Figure 2.4K; Wormbase 2007). By the 16-cell stage, the total mass of mRNA and the number of detected transcripts in each cell negatively correlated with the dynamism of gene regulation in the posterior cells (average R = -0.52 and -0.19) while the length of the cell cycle positively correlated with the dynamism of gene regulation (average R = 0.51, Figure 2.4L). This suggests to us that there are cellular features broadly associated with a cell lineage's progression through the maternal to zygotic transition, including fewer total transcripts and a longer cell cycle.

To quantify the extent to which each cell's transcriptome is unique, we evaluated the number of genes with transcripts detected exclusively in that cell and no others. Again we saw higher numbers of unique transcripts in the non-germ descendants of the P_1 cell. The cell type with the highest number of uniquely expressed genes (176) was the Ex cells (Ea and Ep; Figure 2.4G). These cells have already established an endoderm-specific transcription program (Maduro 2010), and are minutes away from initiating gastrulation by moving from the outside of the embryo to the inside (Nance et al. 2005). These are also the first cells that have a gap phase in their cell cycles, taking 40 minutes to divide compared to ~20 minutes in the other cells of this stage (Edgar & McGhee 1988). Because many of the posterior cells become restricted to a single fate before the anterior cells do, we hypothesized that the posterior cells might express a greater number of cell-specific transcription factors. For each cell type, we calculated the percentage of that cell's unique genes that were transcription factors. We found a larger proportion of mRNAs encoding transcription factors uniquely in the posterior cells (Figure 2.4H), suggesting that these cells are initiating lineage-specific transcriptional programs.

As a cell's transcriptome becomes distinct from that of its neighbors, there are likely several processes involved, including differential transcription, degradation, and segregation of

transcripts during cell division. While all of these processes contribute to the development of an embryo, it is likely that our dataset is most informative regarding transcriptional events. Because of the high false negative rate in scRNA-seq data, cell-specific detection of a transcript is more reliable than cell-specific absence. For this reason we focused most of our following analyses on increases in transcript abundance in specific cells, rather than decreases. In the cases where transcripts of a gene are twice as enriched in a daughter cell compared to its parent, this could be due to transcription or differential enrichment, and our dataset cannot currently distinguish between the two.

The well-documented cell lineage of *C. elegans* tells us the exact lineal relationship between any pair of cells, uniquely allowing us to compare transcriptomes in both space and time. To analyze whether cell-specific features were maintained over time, we generated a correlation matrix comparing the transcriptomes of all cell types to one another (Figure 2.4M). Cells of the 1- and 2-cell stage and all germ cell precursors clustered together with high correlation, indicating that germ cell specific features were common across stages. Otherwise, the strongest correlations were between cells of different lineages but a common stage, suggesting prominent stage-specific expression.

Genes with Spatially Dynamic Expression

When a given transcript is detected across multiple temporal stages in an embryo, the most parsimonious explanation is that the transcript is inherited from parent cells to daughter cells lineally. While we expect some genes to contradict this assumption and be uniquely expressed in cells that are not related by lineage, such a scenario cannot be detected in a whole-embryo timecourse. The present dataset has a high enough resolution both temporally and

spatially that we were able to identify transcripts whose overall expression is continuous throughout consecutive stages, but that are detected in different cell lineages throughout those stages. One such example is *tbx-32* (Figure 2.5A), which was robustly detected in EMS at the 4-cell stage but absent in the daughters of EMS (E and MS) at the following stage. Instead *tbx-32* transcripts appeared in anterior cells ABal, ABar, ABpl and ABpr (also referred to here as ABxx), that are not directly related to EMS by lineage.

To test the validity of this cross-lineage expression pattern, we performed smFISH on intact embryos. We detected *tbx-32* transcripts in EMS at the 4-cell stage and in AB descendants at the 8-cell stage, as our RNA-seq data predicted (Figure 2.5B). *tbx-32* transcripts were more abundant in the 16-cell stage by smFISH than we anticipated from our RNA-seq dataset, but partially degraded transcripts may be more detectable by smFISH (which recognizes many regions of the transcript) than by the RNA-seq method we used (which requires the presence of a polyadenylated tail for detection). This smFISH data allowed us to describe the *tbx-32* expression pattern with an even higher temporal resolution than in the transcriptional lineage. The smFISH data revealed nuclear localization of *tbx-32* transcripts early in the EMS and ABxx cell cycles, and cytoplasmic localization later in these cell cycles. This sequence of localizations suggests that the dynamic pattern of *tbx-32* expression is due to zygotic transcription in these cells. We found five more genes (*tbx-31*, *tbx-40*, Y43D4A.6, Y116A8C.20, ZK666.1; Figure 2.5C) that have patterns similar to *tbx-32*, suggesting that a common mechanism may be regulating all of these genes.

scRNA-seq Data Reveals Synexpressed Sets of Paralogous Genes

The *C. elegans* genome is a snapshot of an evolving document. Continuous duplication and mutation events have produced a genome with many sets of paralogous genes in varying states of divergence. An estimated 32% of *C. elegans* genes have one or more paralogs (Woollard 2005), and we hypothesized that these sets of paralogs are more likely than a random pair of genes to be synexpressed (having transcripts whose expression patterns are highly correlated; Niehrs & Pollet 1999). To test this hypothesis we searched our data for groups of genes that were both synexpressed and similar to one another in sequence.

We found 295 sets of 2-5 genes that were synexpressed and paralogous (Extended Experimental Procedures; Figure 2.6A, Table S3). As a control, we scrambled the gene names in our dataset 100 times and repeated the analysis, finding on average 128 synexpressed paralogous gene sets in these permutations (Figure 2.6B,C). The 295 sets identified using unscrambled data consisted of 640 genes, of which only 126 have a known phenotype (19.7%; WormMine 2016).

scRNA-seq Reveals Genes that are Required for Embryonic Development

To test whether our dataset could lead us to genes that are critical for development but have not yet been appreciated as such, we selected a small group of genes to target by dsRNA injection and test for embryonic lethality. dsRNA injection is more labor-intensive than feeding methods but generally results in more penetrant phenotypes (Ahringer 2006). We selected nine pairs of synexpressed paralogous genes out of the 295 sets identified in Figure 2.6C, and prepared dsRNAs to target each gene. We co-injected each pair into RNAi-hypersensitive *rrf-3* mutant worms. Worms injected with dsRNA targeting one of the nine pairs (T24E12.1 and T24E12.13) produced offspring with 94% embryonic lethality (Figure S4). To test which of

these genes was critical for development we injected dsRNA targeting each of them separately into N2 worms. High levels of embryonic lethality were observed in both conditions (82.7% for T24E12.1 and 70.3% for T24E12.13; Figure 2.6D), suggesting that both these genes are critical for development. Transcripts of these two genes are enriched in AB descendants at the 8- and 16-cell stages (Figure 2.6E). Their mRNA expression patterns were somewhat staggered, with T24E12.1 being more highly detected at earlier stages and in the posterior AB descendants, while T24E12.13 was detected at higher levels in the anterior AB descendants. Knocking down of T24E12.13 by RNAi in previous studies resulted in no detectable phenotype (Kamath et al. 2003; Sönnichsen et al. 2005), while its paralog, T24E12.1, had never previously been tested (Wormbase 2015). This supports our hypothesis that the transcriptional lineage can help identify previously unappreciated genes that are critical for development.

An Interactive Data Visualization Tool to Explore Our Gene Expression Data

To maximize the accessibility of our data, we developed an interactive data visualization tool (available in Chrome and Firefox browsers at http://tintori.bio.unc.edu). With this tool, the user can select which two cells or embryos they wish to compare, and generate a differential gene expression plot that highlights all of the transcripts enriched specifically in either sample (Figure 2.7). All detected transcripts are plotted by their fold change between any two selected samples, and their average expression level. These metrics were chosen because they are less abstracted than p-value and therefore more intuitive, but the user can also filter the data by adjusted p-value using the slider next to the plot.

The interactive tool allows hypothesis-driven analyses (in which the user can query known genes of interest) as well as exploratory analyses (in which the user can discover new

genes of interest). Our scRNA-seq data may be used to explore many fundamental aspects of development, such as specification of distinct cell types such as muscle or intestine, and cell behaviors such as cell cycle control or morphogenesis. We hope our visualization tool will invite researchers working on these and other topics to explore our dataset.

2.3 DISCUSSION

A Transcriptional Lineage to Complement the Completely Defined Cell Lineage of the *C. elegans* Embryo

For decades the *C. elegans* embryo has been a powerful tool for studying cell biology and development, largely because of its invariant cell lineage (Sulston et al. 1983). Here we present a transcriptional lineage that, when paired with the cell lineage, describes the suite of transcripts present in early embryonic cells. Because all cells sampled have a precisely known relationships to each other, this dataset allows a comparison of transcriptomes in space and time, as these cells progressively diverge in fate, morphology, and behavior. As technology improves, scRNA-seq of cells beyond the 16-cell stage will become feasible, ideally allowing the possibility of a transcriptional map for every cell at every stage of development. The challenge of *post hoc* cell identification, explored in a previous study (Hashimshony et al. 2012) and in this manuscript, will continue to be relevant at these later stages of development.

Several research groups have previously performed scRNA-seq on human and mouse cells, and identified their cell types *post hoc* by the transcriptomes (Grün et al. 2015; Yan et al. 2013; Biase et al. 2014; Xue et al. 2013; Zeisel et al. 2015; Jaitin et al. 2014; Trapnell et al. 2014; Satija et al. 2015; Achim et al. 2015; Pollen et al. 2014). The invariant development of *C. elegans* provides a constraint on the possible identities of each transcriptome. This advantage is not present in other systems and can help guide cell-type identification. For example, because

each 4-cell embryo yields exactly 1 transcriptome each of exactly 4 cell types, we know that out of our total of 20 unidentified transcriptomes from this stage (4 cells x 5 replicates), exactly 5 of them are from P_2 cells, 5 are from EMS cells, 5 are from ABa cells, and 5 are from ABp cells. Because we know that a cell from one embryo will have exactly one counterpart in each other embryo, we were able to plot Principle Component Analyses of all transcriptomes, as in Figure 2E,F, and look for clusters containing one cell from each embryo. Given the unique constraints of this system, such clustering suggests that all replicates of a single cell-type are grouping together. The fact that known cell-type markers show transcript enrichment patterns that are consistent with our replicate grouping indicates the accuracy of our gene filtration and iterative PCA approach (Figure 2.2C,G,L,Q).

Another difference between our study and previous scRNA-seq studies that identified cell types *post hoc* is that the *C. elegans* cells used in the present study divide about every 20 minutes, whereas the human and mouse cells of previous studies divide approximately every 24 hours. Given this comparatively short cell cycle, it is remarkable that the posterior *C. elegans* cells have such distinct transcript signatures, and by the same token perhaps not surprising that the anterior cells are difficult to distinguish.

Previous studies have described transcriptomes at these stages of development at a lower spatial resolution (Baugh et al. 2003; Baugh et al. 2005; Hashimshony et al. 2012; Hashimshony et al. 2015). By leaving embryos intact until immediately before sample collection, sequencing every individual cell at each stage, and using technology that captures full-length mRNAs, we have expanded upon these previous datasets. Our method preserves fate-determining cell signaling events, allows for comparisons between groups of cells that were never sequenced

separately before (such as AB descendants) and allows for inquiry into cell-specific variation in transcript splicing.

Cells of the Early Embryo Can Be Identified by Their Transcriptomes Alone

We have assigned a cell identity to each transcriptome based on its transcript abundance data, cross-referenced to known expression patterns and *in situ* RNA hybridization. The transcriptomes of some cell types (particularly the P_1 descendants) grouped together tightly, were clearly distinct from other cell types, and had identities confirmed by well-studied genes, making us confident in our assessments. For the anterior cell types whose transcriptomes were less distinct from one another, we have a lower confidence in our assignments (as in Figure 2.3D′). We consider this paucity of distinguishing features to be an interesting biological result, suggesting that it would make little difference if transcriptomes identities were mis-assigned between these cell-types. The current understanding of these cells' developmental potential supports the notion that they should be difficult to distinguish from each other. For example, the sister cells ABa and ABp of the 4-cell stage are initially developmentally equivalent, and the differences between them are not established until after cytokinesis separates them (Priess & Thomson 1987). In the future, if features are identified that more clearly differentiate these cell types, our existing single-cell transcriptomes can be revisited with those features in mind.

Previous studies that have measured transcript abundance in cells of the early embryo have either measured whole-embryo transcript levels (Baugh et al. 2003; Baugh et al. 2005; Levin et al. 2012), or measured only parts of the embryo at a single-cell resolution and the rest of the embryo in clusters of related cells (Hashimshony et al. 2015). These clusters of cells were sampled by dissecting embryos starting at the 2-cell stage and allowing the isolated cells to

divide in culture, then sequencing the group of descendants. This allowed descendants of founder cells to be harvested at later time points than in our study, but kept the cells naïve to critical signaling events that take place in intact embryos. With our dataset, by leaving all cells intact in the embryo until minutes before sampling, we captured single-cell transcriptomes while allowing the cell-cell signaling necessary for proper development to occur, and we detected the transcriptional results of this signaling (Figure 2.3B,F).

A Stark Contrast in mRNA Composition Between Germ Cell Precursors and Somatic Cells

One pattern that is apparent when comparing gene expression across all cell types (Figure 2.4B) is that there is a prevalent distinction between the mRNA composition of the somatic cells and the germ cells (including the somatic sister of each germ cell precursor). Previous studies, such as Seydoux & Fire 1994, have observed this contrast in transcript composition between the germ and soma. Their reliance on *in situ* hybridization necessarily restricted the number of such genes they were able to study (10 genes), whereas the present genome-wide study expands their findings to thousands of genes. Differences between "immortal" germ cells and "mortal" somatic cells have fascinated researchers for over a century (Weismann 1893; Boveri 1910; Schierenberg & Strome 1992; Lai & King 2013; Lehmann & Ephrussi 2007; Yamanaka 2007). The present dataset quantifies this dichotomy and the transition from one state to the other over time. This dataset includes before, during, and after snapshots of somatic descendants of germ cell precursors, in their transition from the germ-like profiles of their parent cell to the somatic profiles of their descendants.

Cross-lineage Expression Patterns Highlight Genes that May Share Mechanisms of Gene Regulation

tbx-32 and the five other genes with similar expression patterns are examples of genes whose expression is not continuous from parent to daughter cell, but rather appears in one cell type (EMS) at one stage, then in a different lineage of cells (ABxx) at the next stage. The EMS cell at the 4-cell stage and one of these ABxx cells (ABar) at the 8-cell stage have another feature in common, which is that both orient their mitotic spindles in response to Wnt signaling (Walston et al. 2004).The fact that this specific expression pattern is shared by several genes suggests that a common mechanism may be regulating all of these genes, possibly the previously characterized Wnt signaling. Alternatively, these six genes may play a role in establishing which cells are capable of responding to Wnt signaling.

Identifying Critical Regulators of Development

Testing a small subset of genes, we identified two that are critical for embryonic development (Figure 2.6D). This indicates that our dataset may be well-suited to highlight previously unappreciated key regulators. These two genes are similar in sequence, and have similar but slightly staggered transcript enrichment patterns (Figure 2.6E). The staggering of these two patterns may represent subfunctionalization after a gene duplication event. This observation suggests that by considering both homology and spatiotemporal transcript abundance, our dataset may reveal patterns about divergence in sequence and function after a gene duplication event.

Although we saw embryonic lethal phenotypes in only 2 of the 18 genes we tested by dsRNA injection, we expect that a higher proportion of the genes highlighted by our dataset are

likely to be important, for example for embryonic functions not required for hatching, for postembryonic development, or for stress tolerance.

2.4 MATERIALS AND METHODS

Worm Husbandry and Embryo Dissections

All worms were grown at 20°C and dissected at room temperature (21-24°C). Single embryos were selected at 10-20 minutes before the desired stage and dissected as in Edgar & Goldstein 2012 (modifications described in Extended Experimental Procedures).

RNA Preparation, Sequencing, and RPKM Generation

cDNA was generated using the SMARTer Ultra Low RNA Input for Illumina Sequencing Kit, and sequencing libraries were prepared using the Nextera XT kit, both according to manufacturers instructions. Sequences from this study are available at NCBI GEO GSE77944 (reviewer access: http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?token=uhqtwmumzvshroj&acc=GSE77944).

RNAi

RNAs were combined and diluted to a total of 1 ug/uL for each condition. 15-22 young adult worms were injected for each condition. *rrf-3* mutant worms (PK1429) were used for experiments in Figure S4, and N2 worms were used for experiments in Figure 2.6D. Embryonic lethality was calculated as the percent of unhatched embryos remaining 24 hours after mothers were removed from the plate, out of the total unhatched and hatched progeny.

Single Molecule Fluorescent *in situ* Hybridization

N2 (Figure 2.3) or LP306 worms (containing a GFP membrane marker, Figure 2.5) were grown at 20°C and embryos were prepared as in Shaffer et al. 2013 and Ji & van Oudenaarden 2012 (details in Extended Experimental Procedures).

2.5 ACKNOWLEDGEMENTS

The authors would like to thank Corbin Jones and the UNC High Throughput Sequencing Facility for their generosity during scRNA-seq protocol optimization. The authors declare no conflicts of interest. BG and SCT were supported by NIH grant R01 GM083071. JDL, EON and SCT were supported by NIH grant R01 GM104050. SCT was supported by NIH grant F31 HD088128 and by a NSF GRFP award.

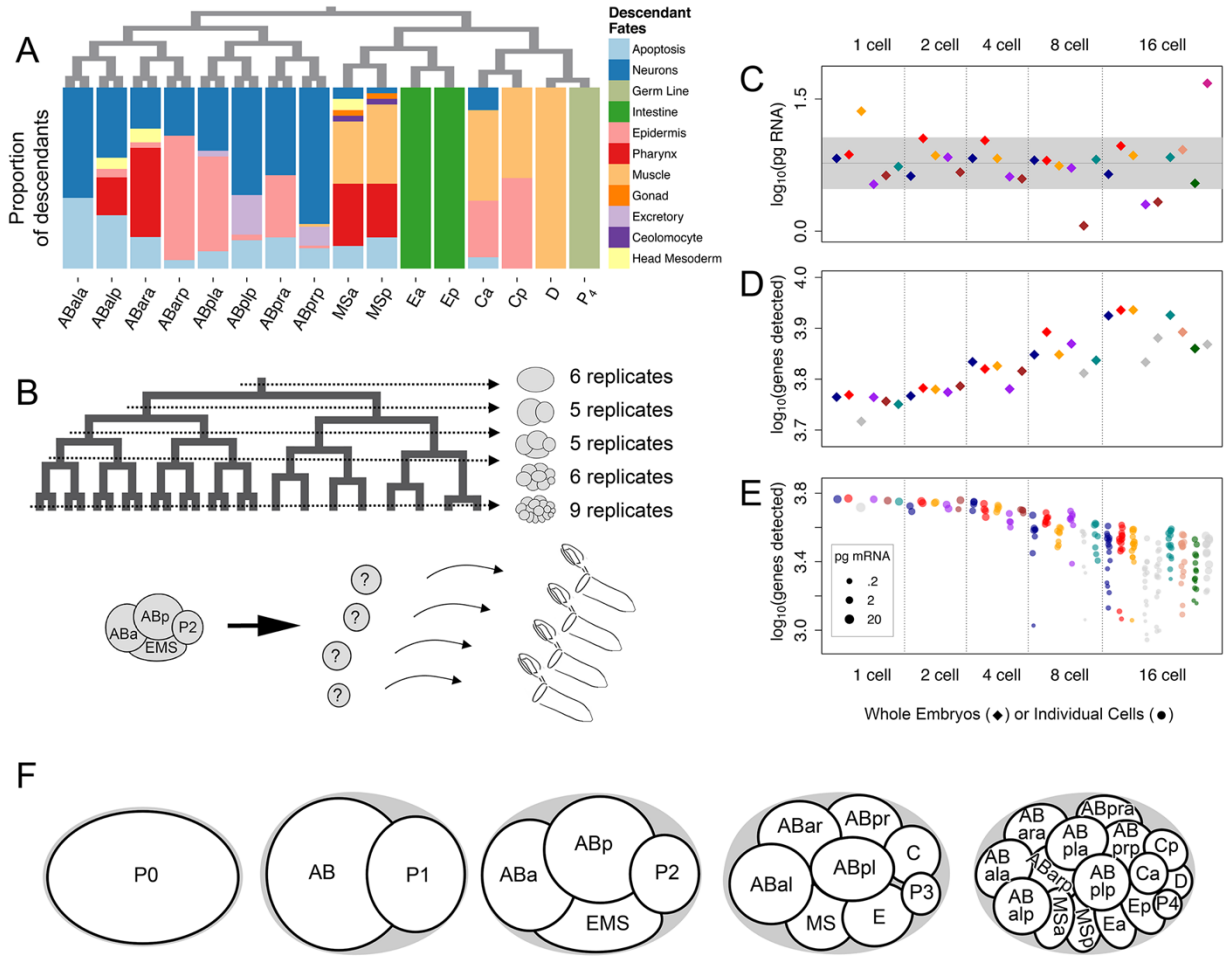

Figure 2.1: Single-cell mRNA-seq libraries for complete sets of cells from *C. elegans* embryos of the 1-, 2-, 4-, 8- and 16-cell stages.

(A) Terminal cell fates of descendants of each cell of the 16-cell embryo. Terminal fates were calculated from Sulston et al. 1983, and refer to cell fates at the time of the first larval hatching.

(B) Schematic of samples that were hand-dissected and prepared for scRNA-seq. The 4-cell stage is diagrammed below for illustration.

(C) The total mass of mRNA detected from each embryo (diamonds). Embryos whose total mass of mRNA differed from the average by more than one standard deviation (plotted outside of gray band) were excluded from subsequent analyses.

(D) The number of genes whose transcripts were detected in each whole embryo (diamonds).

(E) The number of genes whose transcripts were detected in each individual cell (circle).

(F) Key of the names of each cell from the zygote to the 16-cell stage.

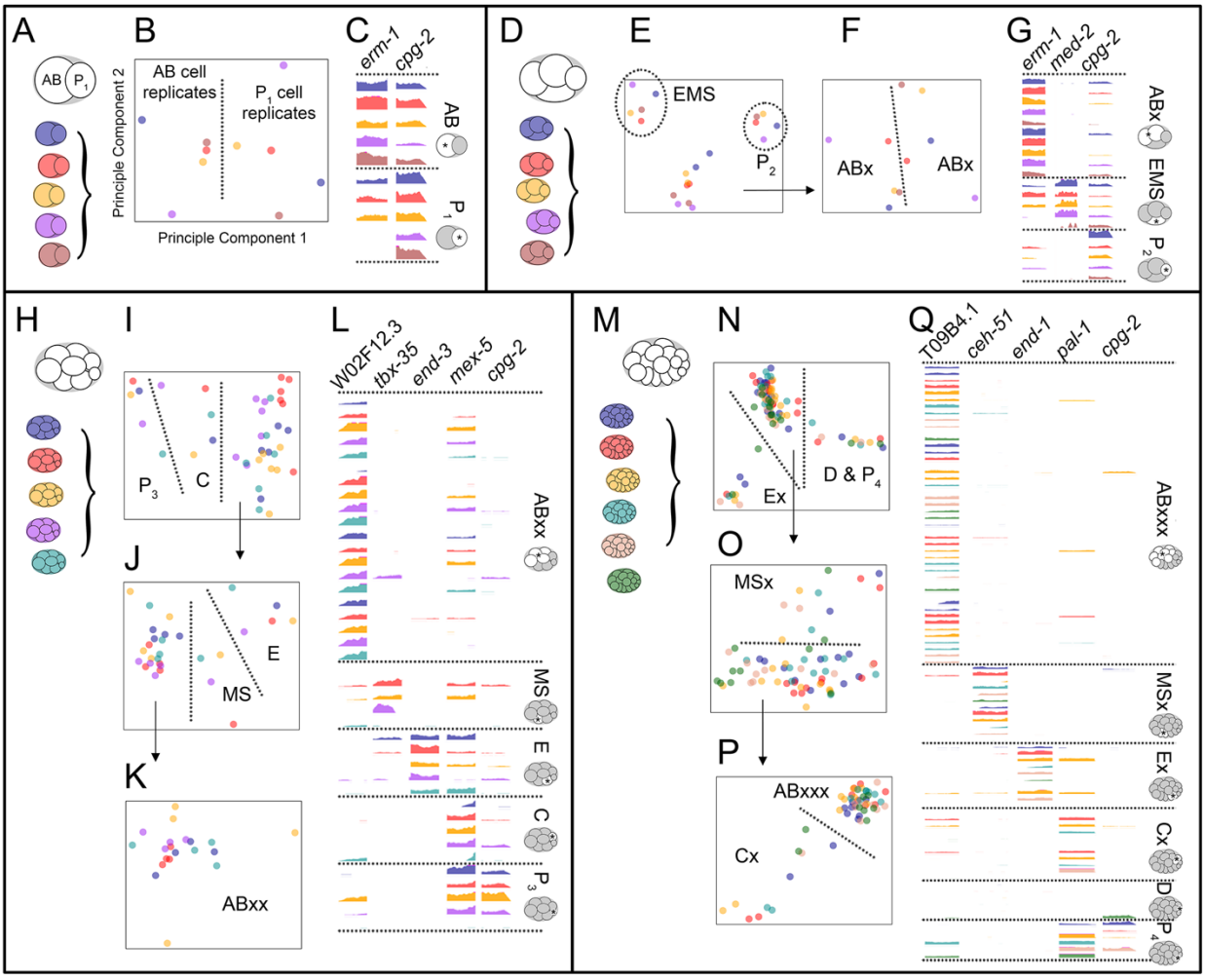

Figure 2.2: Replicates of each cell type were grouped by transcript signatures and identified by candidate gene expression.

(A-C) Transcriptomes of cells from the 2-cell stage (A) were subjected to Principle Component Analysis (PCA) (B) using only data from reproducibly differentially enriched genes, as selected by our algorithm (details in Experimental Procedures). (C) Genome browser tracks of the last exon of *erm-1* (AB-enriched) and *cpg-2* (P_1-enriched). Colors correspond to embryo of origin. Heights of tracks indicate read count density. All y-axes of genome browser tracks are scaled consistently within each panel.

(D-G) Transcriptomes of cells from the 4-cell stage (D) were subjected to PCA (E). (F) PCA of the 10 transcriptomes that were not resolved in (E). (G) Genome browser tracks of the last exon of *erm-1* (AB-enriched), *med-2* (EMS-enriched) and *cpg-2* (P_2-enriched).

(H-L) Transcriptomes of cells from the 8-cell stage (H) were subjected to PCA using iteratively generated sets of informative genes (I-K). (L) Genome browser tracks of the last exon of W02F12.3 (ABxx-enriched), *tbx-35* (MS-enriched), *end-3* (E-enriched), *mex-5* (C- and P_3-enriched), and *cpg-2* (P_3-enriched).

(M-Q) Transcriptomes of cells from the 16-cell stage (M) were subjected to PCA using iteratively generated sets of informative genes (N-P). (Q) Genome browser tracks of the last exon of T09B4.1 (ABxxx-specific), *ceh-51* (MSx-specific), *end-1* (Ex-specific), *pal-1* (Cx- and P_4-specific), and *cpg-2* (P_4-specific).

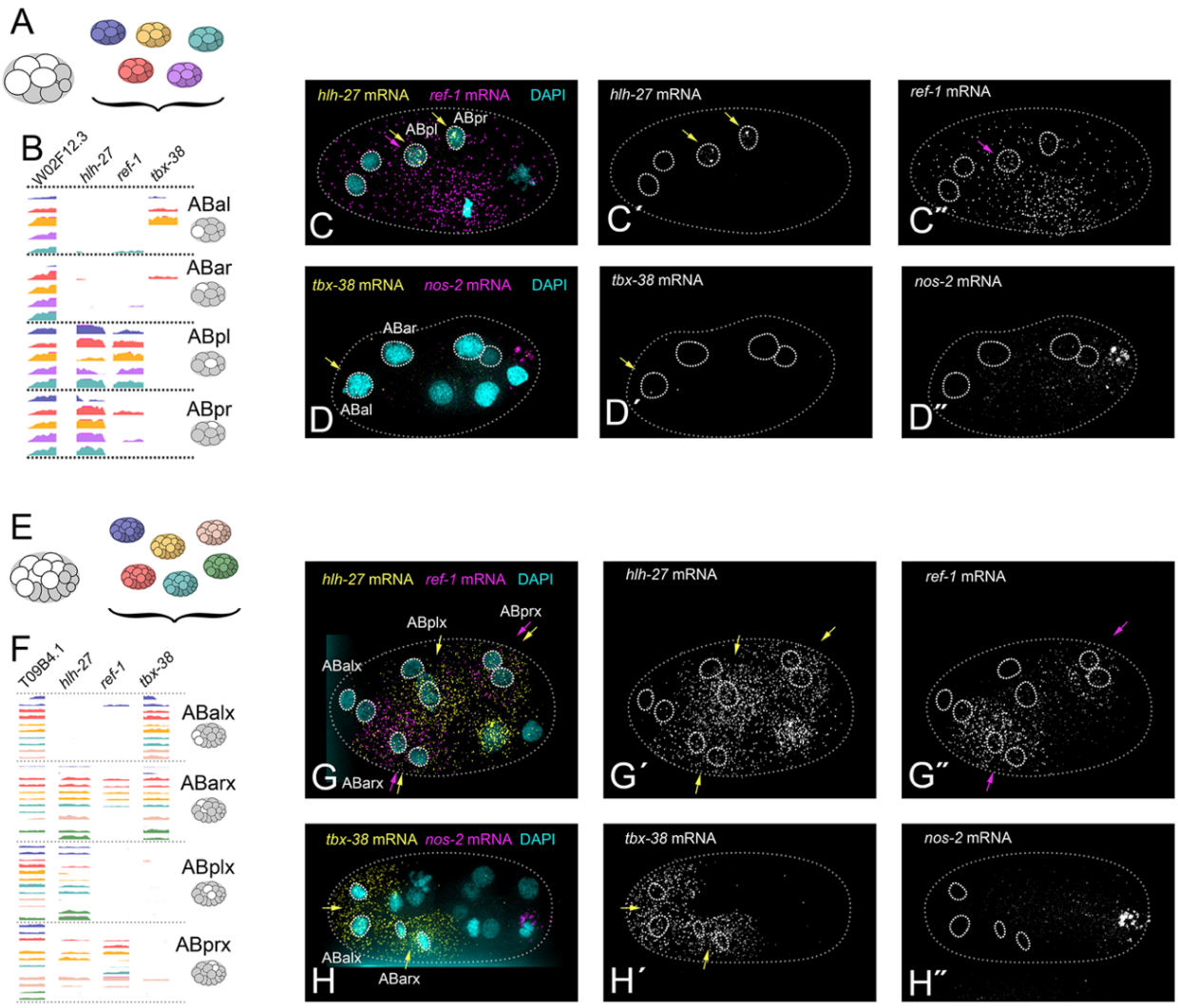

Figure 2.3: Differential transcript enrichment of notch target genes in cells that could not be distinguished by global transcript signatures.
(A) AB descendants from five replicates of the 8-cell stage embryo.
(B) Genome browser tracks of ABxx transcriptomes, sorted into groups based on expression of notch target genes *hlh-27*, *ref-1* and *tbx-38* (Extended Methods). Last exons only are shown.
(C) Example of smFISH targeting *hlh-27* (C′, yellow arrows) and *ref-1* (C″, purple arrows) transcripts in intact 6- or 8-cell stage embryos (*hlh-27* pattern seen in 100% of embryos, n=4. *ref-1* pattern seen in 75% of embryos, n=4. Remaining embryo showed ubiquitous *ref-1* staining).
(D) Example of smFISH targeting *tbx-38* (D′, yellow arrows) in intact 8-cell stage embryos (pattern seen in 33% of embryos, n=3. 67% of embryos showed equal *tbx-38* expression in ABal and ABar). (D″) *nos-2* (P_3-specific) marks the posterior of the embryo.
(E) AB descendants from six replicates of the 16-cell stage embryo.
(F) Genome browser tracks of ABxxx transcriptomes, sorted into four groups based on a PCA using only notch target gene expression (shown in Figure S2D). Last exons only are shown.

(G) Example of smFISH targeting *hlh-27* (G′, yellow arrows) and *ref-1* (G″, purple arrows) transcripts in intact 15-cell stage embryos (both patterns seen in 100% of embryos, *hlh-27* n=5, *ref-1* n=2).
(H) Example of smFISH targeting *tbx-38* (H′, yellow arrows) in intact 15-cell stage embryos (pattern seen in 100% of embryos, n=14). *nos-2* (P_4-specific) marks the posterior of the embryo.

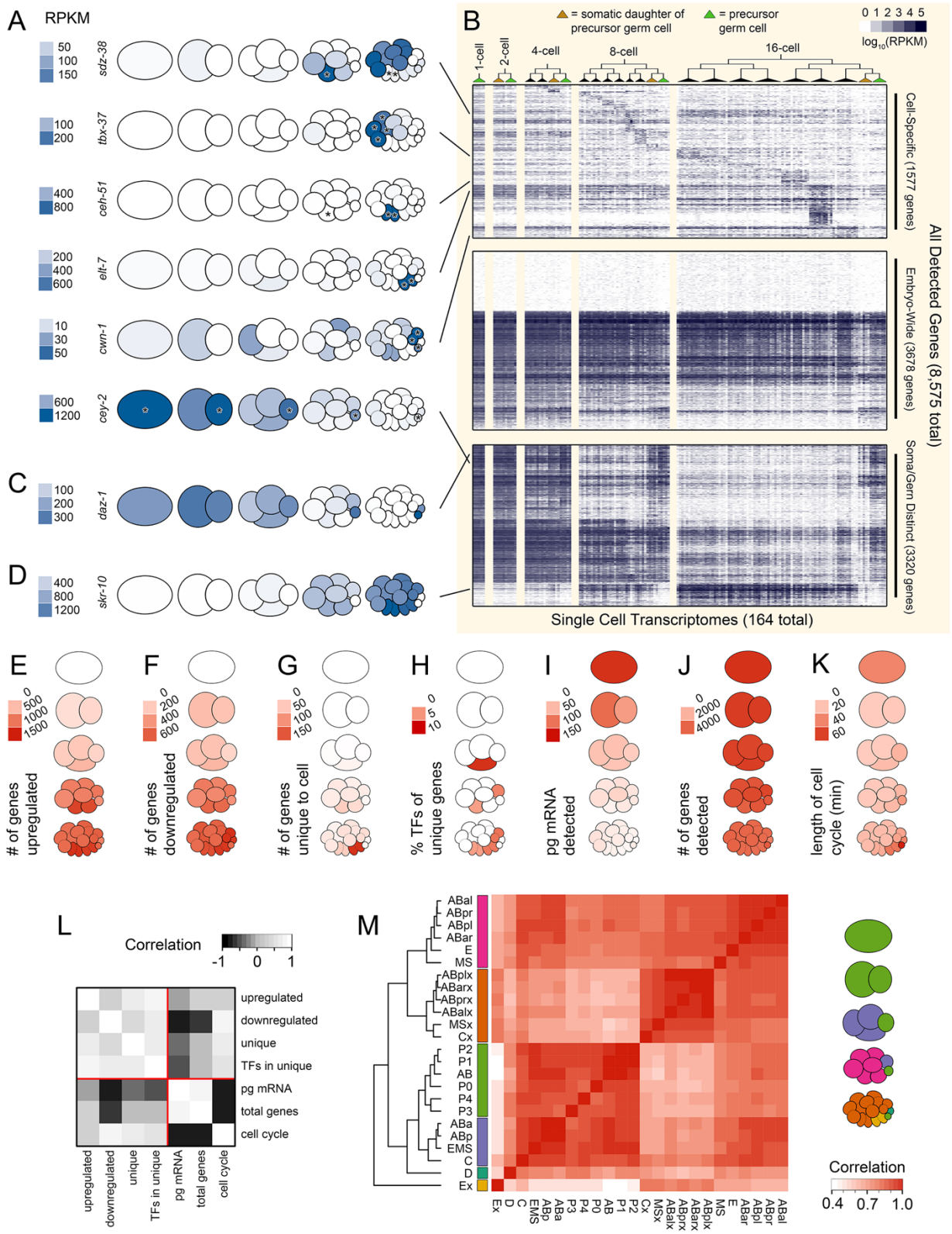
A
RPKM
sdz-38
tbx-37
ceh-51
elt-7
cwn-1
cey-2
B
= somatic daughter of precursor germ cell
= precursor germ cell
log10(RPKM)
1-cell
2-cell
4-cell
8-cell
16-cell
Cell-Specific (1577 genes)
Embryo-Wide (3678 genes)
Soma/Germ Distinct (3320 genes)
All Detected Genes (8,575 total)
Single Cell Transcriptomes (164 total)
C
daz-1
D
skr-10
E
# of genes upregulated
F
# of genes downregulated
G
# of genes unique to cell
H
% TFs of unique genes
I
pg mRNA detected
J
# of genes detected
K
length of cell cycle (min)
L
Correlation
upregulated
downregulated
unique
TFs in unique
pg mRNA
total genes
cell cycle
M
ABal
ABpr
ABpl
ABar
E
MS
ABplx
ABarx
ABprx
ABalx
MSx
Cx
P2
P1
AB
P0
P4
P3
ABa
ABp
EMS
C
D
Ex
Correlation
0.4 0.7 1.0

Figure 2.4: Differential activation of the zygotic genome in each cell lineage.
(A) Transcript abundances of six genes with previously known expression patterns, heat-mapped on to pictograms of the embryo (key in Figure 2.1F). Asterisks indicate the cells in which we expected expression, based on the literature; *sdz-38* expected in E, Ex (Ea and Ep); *tbx-37* expected in ABalx (ABala and ABalp), ABarx (ABara and ABarp); *ceh-51* expected in MS, MSx (MSa and MSp); *elt-7* expected in Ex (Ea and Ep); *cwn-1* expected in Cx (Ca and Cp), D; *cey-2* expected in P_0, P_1, P_2, P_3, P_4 (references in Main Text).
(B) Heatmap of transcript abundances of all 8,575 present genes (y-axis) in each cell throughout time and space (x-axis). Only transcriptomes that passed quality filtration were plotted (164 out of 219). The y-axis along the top third of the heatmap is scaled twice as large as the bottom two thirds, to show detail.
(C) Transcript abundance data for *daz-1* (a maternally inherited gene required for meiosis; Karashima et al. 2000), an example of a transcript we detected in only the germ cells and their sister cells.
(D) Transcript abundance data for *skr-10* (a member of the ubiquitin ligase complex; Yamanaka et al. 2002), an example of a transcript we detected in only somatic cells.
(E) The number of upregulated genes for each cell type. Genes were scored as upregulated in a cell if their transcripts were at least twice as abundant as in any ancestors of that cell.
(F) The number of downregulated genes for each cell type. Genes were scores as downregulated in a cell if their transcript abundances were half or less that of an ancestor.
(G) The number of cell-specific, or unique, genes. Genes were scored as unique to a cell type if their transcript abundance was at least 10 times higher than in any other cell type in the dataset.
(H) Percentage of each cell type's unique genes, as defined in (G), that are transcription factors.
(I) Mass of mRNA per cell as calculated using concentrations of control mRNA spike-ins.
(J) Number of genes detected above 25 RPKM in each cell.
(K) Length of cell cycle for each cell.
(L) Pearson correlation of E-K across all cell types (excluding germ cell precursors, which are known to be transcriptionally distinct; Schaner & Kelly 2006).
(M) Matrix of the correlation coefficients of all cell types' transcriptomes. Six branches of highly correlated cell types are color coded in the cartoon to the right.

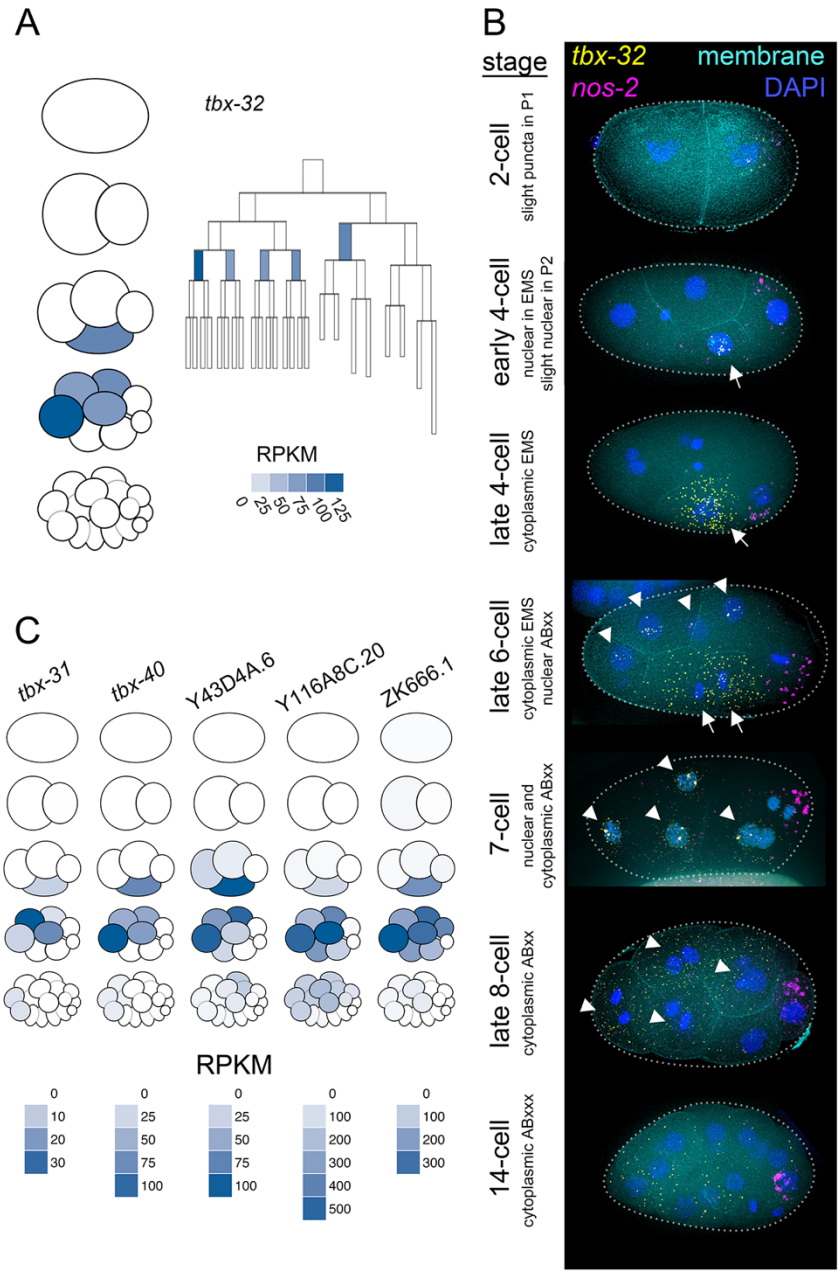

Figure 2.5: Spatially dynamic gene expression is revealed by high resolution data.
(A) A cell lineage map and a pictogram of the 1- through 16-cell stages. Color corresponds to transcript abundance data for *tbx-32* in each cell type.
(B) smFISH of *tbx-32*. 100% of 2-cell stage embryos (n=2), 83% of 4-cell stage embryos (n=6, one embryo showed ubiquitous staining), 100% of 6- to 8-cell stage embryos (n=6), and 100% of 12- to 15-cell stage embryos (n=3) showed this pattern.
(C) Pictograms for 5 genes showing transcript enrichement patterns similar to that of *tbx-32*.

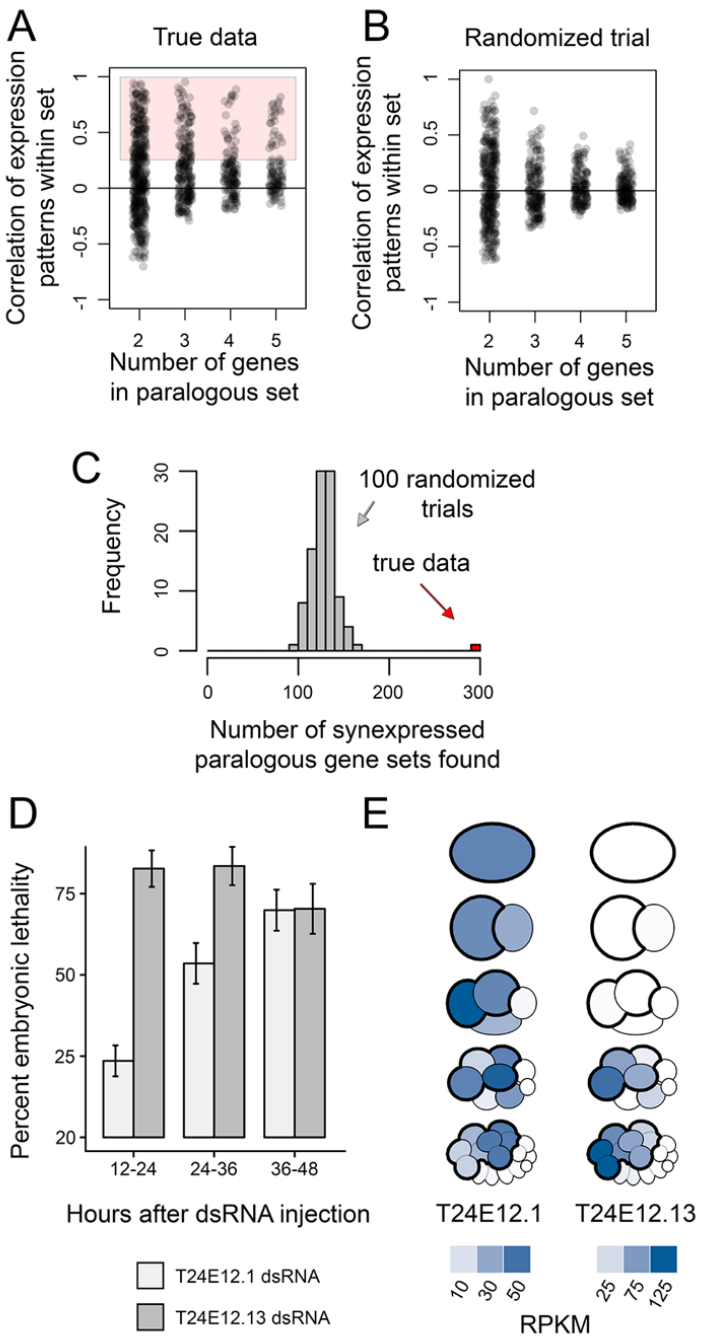

Figure 2.6: Previously unappreciated paralogous, synexpressed genes are critical for development.
(A) Correlations of expression patterns for sets of 2-5 genes that are similar to each other in sequence. 295 sets of genes had a correlation coefficient greater than 0.25, and were considered paralogous and synexpressed.
(B) Fewer sets of paralagous genes are highly correlated in a scrambled dataset.
(C) Histogram of the number of synexpressed paralogous gene sets detected in our dataset (red bar) and in 100 datasets randomized by scrambling gene names without replacement (gray bars).

(D,E) MA plots comparing gene expression in the AB descendants to gene expression in the other cells of the 8-cell (D) or 16-cell (E) stage. One pair of paralogous genes (T24E12.1 and T24E12.13) is highlighted in cyan.
(F) Pictograms showing quantitative transcript abundance data for the genes highlighted in (D,E).
(G) Lethality phenotype observed in embryos in which T24E12.1 and T24E12.13 were targeted by co-injection of dsRNA.

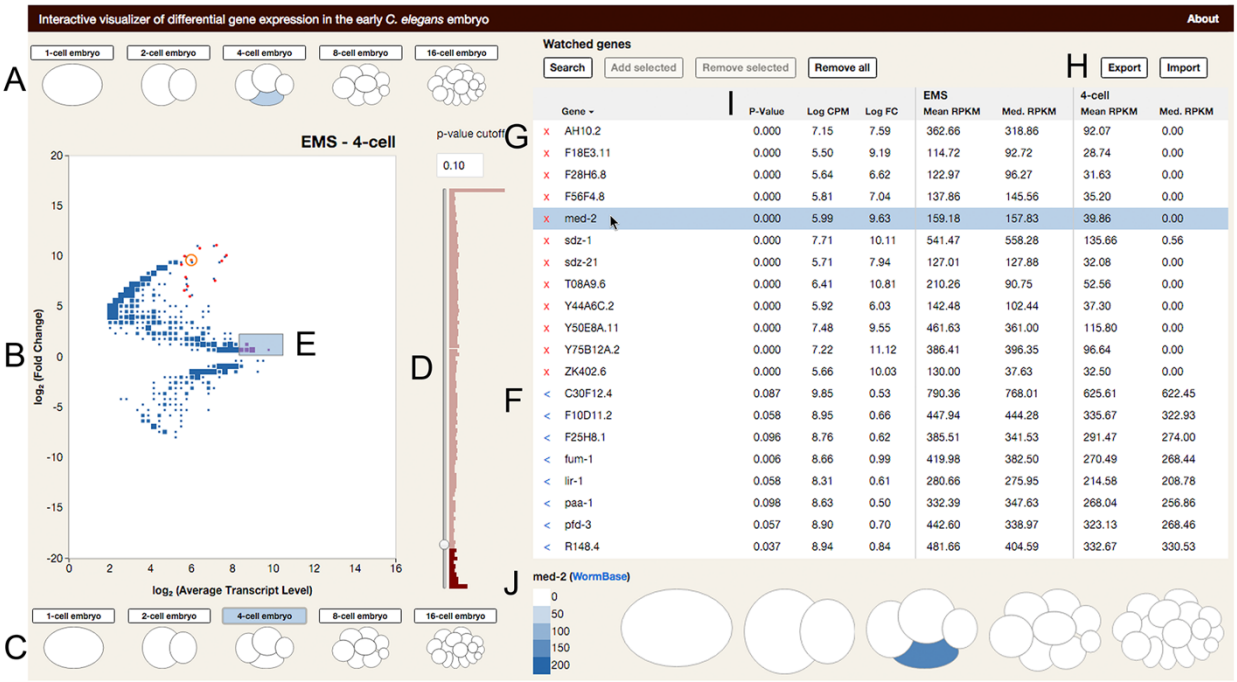

Figure 2.7: An interactive data visualization tool for querying the transcriptional lineage
Still image of data visualization tool. Full version available in Chrome and Firefox broswers at http://tintori.bio.unc.edu.
(A-C) Sample selection. The user clicks on the cells or whole embryos they wish to compare on the top (A) and bottom (C) of the plot. When a new sample is selected, the plot (B) is redrawn to reflect the selected comparison. Size of points in B scales to the number of genes represented by each dot.
(D-H) Gene selection. (D) The user can filter genes by adjusted P-value of differential enrichment between samples. (E) Clicking on a point or selecting a swath of points on the plot adds genes and their data to the Selected Genes table (F). Known genes can be added directly, by typing their names into the search bar. (G) The Watched Genes table is curated by adding Selected Genes individually or in bulk. (H) The Watched Genes table can be exported, and lists of genes can be imported to the Watched Genes table in bulk.
(I-J) Gene expression metrics. (I) The gene tables are sortable by name, average expression level, fold change, significance of differential enrichment, and expression levels in either sample being compared. (J) Clicking on a gene in the table reveals a cartoon of the embryo over all five stages. Each cell is colored corresponding to the transcript level of the highlighted gene.

REFERENCES

Achim, K., Pettit, J.-B., Saraiva, L. R., Gavriouchkina, D., Larsson, T., Arendt, D., & Marioni, J. C. (2015). High-throughput spatial mapping of single-cell RNA-seq data to tissue of origin. *Nature Biotechnology*, *33*(5), 503–509. http://doi.org/10.1038/nbt.3209

Anders, S., Pyl, P. T., & Huber, W. (2014). HTSeq—a Python framework to work with high-throughput sequencing data. *Bioinformatics*, *31*(2), btu638–169. http://doi.org/10.1093/bioinformatics/btu638

Ahringer, J. (2006). Reverse Genetics. *WormBook : the Online Review of C. Elegans Biology*, 1–6. http://doi.org/10.1895/wormbook.1.47.1

Baker, S. C., Bauer, S. R., Beyer, R. P., Brenton, J. D., Bromley, B., Burrill, J., et al. (2005). The External RNA Controls Consortium: a progress report. *Nature Methods*, *2*(10), 731–734. http://doi.org/10.1038/nmeth1005-731

Baugh, L. R., Hill, A. A., Claggett, J. M., Hill-Harfe, K., Wen, J. C., Slonim, D. K., et al. (2005). The homeodomain protein PAL-1 specifies a lineage-specific regulatory network in the C. elegans embryo. *Development (Cambridge, England)*, *132*(8), 1843–1854. http://doi.org/10.1242/dev.01782

Baugh, L. R., Hill, A. A., Slonim, D. K., Brown, E. L., & Hunter, C. P. (2003). Composition and dynamics of the Caenorhabditis elegans early embryonic transcriptome. *Development (Cambridge, England)*, *130*(5), 889–900.

Biase, F. H., Cao, X., & Zhong, S. (2014). Cell fate inclination within 2-cell and 4-cell mouse embryos revealed by single-cell RNA sequencing. *Genome Research*, *24*(11), 1787–1796. http://doi.org/10.1101/gr.177725.114

Björklund, Å. K., Forkel, M., Picelli, S., Konya, V., Theorell, J., Friberg, D., et al. (2016). The heterogeneity of human CD127(+) innate lymphoid cells revealed by single-cell RNA sequencing. *Nature Immunology*. http://doi.org/10.1038/ni.3368

Boveri, T. (1910). Über die Teilung centrifugierter Eier von Ascaris megalocephala. *Archiv Für Entwicklungsmechanik Der Organismen*, *30*(2), 101–125. http://doi.org/10.1007/BF02263806

Bowerman, B., Ingram, M. K., & Hunter, C. P. (1997). The maternal par genes and the segregation of cell fate specification activities in early Caenorhabditis elegans embryos. *Development (Cambridge, England)*, *124*(19), 3815–3826.

Broitman-Maduro, G., Lin, K. T.-H., Hung, W. W. K., & Maduro, M. F. (2006). Specification of the C. elegans MS blastomere by the T-box factor TBX-35. *Development (Cambridge, England)*, *133*(16), 3097–3106. http://doi.org/10.1242/dev.02475

Broitman-Maduro, G., Owraghi, M., Hung, W. W. K., Kuntz, S., Sternberg, P. W., & Maduro, M. F. (2009). The NK-2 class homeodomain factor CEH-51 and the T-box factor TBX-35 have overlapping function in C. elegans mesoderm development. *Development (Cambridge, England)*,

136(16), 2735–2746. http://doi.org/10.1242/dev.038307

Cheung, T. H., & Rando, T. A. (2013). Molecular regulation of stem cell quiescence. *Nature Reviews. Molecular Cell Biology*, *14*(6), 329–340. http://doi.org/10.1038/nrm3591

Deshpande, G., Calhoun, G., & Schedl, P. (2004). Overlapping mechanisms function to establish transcriptional quiescence in the embryonic Drosophila germline. *Development (Cambridge, England)*, *131*(6), 1247–1257. http://doi.org/10.1242/dev.01004

Edgar, L. G., & Goldstein, B. (2012). Culture and Manipulation of Embryonic Cells. *Methods in Cell Biology*, *107*, 151–175. http://doi.org/10.1016/B978-0-12-394620-1.00005-9

Edgar, L. G., & McGhee, J. D. (1988). DNA synthesis and the control of embryonic gene expression in C. elegans. *Cell*, *53*(4), 589–599.

Edgar, L. G., Wolf, N., & Wood, W. B. (1994). Early transcription in Caenorhabditis elegans embryos. *Development (Cambridge, England)*, *120*(2), 443–451.

Gleason, J. E., Szyleyko, E. A., & Eisenmann, D. M. (2006). Multiple redundant Wnt signaling components function in two processes during C. elegans vulval development. *Developmental Biology*, *298*(2), 442–457. http://doi.org/10.1016/j.ydbio.2006.06.050

Gordon, A., & Hannon, G. J. (2010). FASTX Toolkit. http://hannonlab.cshl.edu/fastx_toolkit/index.html .

Grün, D., Lyubimova, A., Kester, L., Wiebrands, K., Basak, O., Sasaki, N., et al. (2015). Single-cell messenger RNA sequencing reveals rare intestinal cell types. *Nature*, *525*(7568), 251–255. http://doi.org/10.1038/nature14966

Hashimshony, T., Feder, M., Levin, M., Hall, B. K., & Yanai, I. (2015). Spatiotemporal transcriptomics reveals the evolutionary history of the endoderm germ layer. *Nature*, *519*(7542), 219–222. http://doi.org/10.1038/nature13996

Hashimshony, T., Wagner, F., Sher, N., & Yanai, I. (2012). CEL-Seq: single-cell RNA-Seq by multiplexed linear amplification. *Cell Reports*, *2*(3), 666–673. http://doi.org/10.1016/j.celrep.2012.08.003

Jaitin, D. A., Kenigsberg, E., Keren-Shaul, H., Elefant, N., Paul, F., Zaretsky, I., et al. (2014). Massively Parallel Single-Cell RNA-Seq for Marker-Free Decomposition of Tissues into Cell Types. *Science (New York, N.Y.)*, *343*(6172), 776–779. http://doi.org/10.1126/science.1247651

Ji, N., & van Oudenaarden, A. (2012). Single molecule fluorescent in situ hybridization (smFISH) of C. elegans worms and embryos. *WormBook : the Online Review of C. Elegans Biology*, 1–16. http://doi.org/10.1895/wormbook.1.153.1

Kamath, R. S., Fraser, A. G., Dong, Y., Poulin, G., Durbin, R., Gotta, M., et al. (2003). Systematic functional analysis of the Caenorhabditis elegans genome using RNAi. *Nature*, *421*(6920), 231–237. http://doi.org/10.1038/nature01278

Karashima, T., Sugimoto, A., & Yamamoto, M. (2000). Caenorhabditis elegans homologue of

the human azoospermia factor DAZ is required for oogenesis but not for spermatogenesis. *Development (Cambridge, England)*, *127*(5), 1069–1079.

Kim, D., Pertea, G., Trapnell, C., Pimentel, H., & Kelley, R. (2013). TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biology, 14*(4), http://doi.org/10.1186/gb-2013-14-4-r36

Lai, F., & King, M. L. (2013). Repressive translational control in germ cells. *Molecular Reproduction and Development*, *80*(8), 665–676. http://doi.org/10.1002/mrd.22161

Lehmann, R., & Ephrussi, A. (2007). Germ Plasm Formation and Germ Cell Determination in Drosophila. In *Ciba Foundation Symposium 182 - Germline Development* (pp. 282–304). Chichester, UK: John Wiley & Sons, Ltd. http://doi.org/10.1002/9780470514573.ch16

Levin, M., Hashimshony, T., Wagner, F., & Yanai, I. (2012). Developmental Milestones Punctuate Gene Expression in the Caenorhabditis Embryo. *Developmental Cell*, *22*(5), 1101–1108. http://doi.org/10.1016/j.devcel.2012.04.004

Maduro, M. F. (2010). Cell fate specification in the C. elegans embryo. *Developmental Dynamics : an Official Publication of the American Association of Anatomists*, *239*(5), 1315–1329. http://doi.org/10.1002/dvdy.22233

Maduro, M. F., Broitman-Maduro, G., Mengarelli, I., & Rothman, J. H. (2007). Maternal deployment of the embryonic SKN-1-->MED-1,2 cell specification pathway in C. elegans. *Developmental Biology*, *301*(2), 590–601. http://doi.org/10.1016/j.ydbio.2006.08.029

Maduro, M. F., Hill, R. J., Heid, P. J., Newman-Smith, E. D., Zhu, J., Priess, J. R., & Rothman, J. H. (2005). Genetic redundancy in endoderm specification within the genus Caenorhabditis. *Developmental Biology*, *284*(2), 509–522. http://doi.org/10.1016/j.ydbio.2005.05.016

Maduro, M. F., Meneghini, M. D., Bowerman, B., Broitman-Maduro, G., & Rothman, J. H. (2001). Restriction of mesendoderm to a single blastomere by the combined action of SKN-1 and a GSK-3beta homolog is mediated by MED-1 and -2 in C. elegans. *Molecular Cell*, *7*(3), 475–485.

Mello, C. C., Schubert, C., Draper, B., Zhang, W., Lobel, R., & Priess, J. R. (1996). The PIE-1 protein and germline specification in C. elegans embryos. *Nature*, *382*(6593), 710–712. http://doi.org/10.1038/382710a0

Nance, J., Lee, J.-Y., & Goldstein, B. (2005). Gastrulation in C. elegans. *WormBook : the Online Review of C. Elegans Biology*, 1–13. http://doi.org/10.1895/wormbook.1.23.1

Neves, A., & Priess, J. R. (2005). The REF-1 family of bHLH transcription factors pattern C. elegans embryos through Notch-dependent and Notch-independent pathways. *Developmental Cell*, *8*(6), 867–879. http://doi.org/10.1016/j.devcel.2005.03.012

Niehrs, C., & Pollet, N. (1999). Synexpression groups in eukaryotes. *Nature*, *402*(6761), 483–487. http://doi.org/10.1038/990025

Osborne Nishimura, E., Zhang, J. C., Werts, A. D., Goldstein, B., & Lieb, J. D. (2015).

Asymmetric transcript discovery by RNA-seq in C. elegans blastomeres identifies neg-1, a gene important for anterior morphogenesis. *PLoS Genetics*, *11*(4), e1005117. http://doi.org/10.1371/journal.pgen.1005117

Pollen, A. A., Nowakowski, T. J., Shuga, J., Wang, X., Leyrat, A. A., Lui, J. H., et al. (2014). Low-coverage single-cell mRNA sequencing reveals cellular heterogeneity and activated signaling pathways in developing cerebral cortex. *Nature Biotechnology*, *32*(10), 1053–1058. http://doi.org/10.1038/nbt.2967

Priess, J. R. (2005). Notch signaling in the C. elegans embryo. *WormBook : the Online Review of C. Elegans Biology*, 1–16. http://doi.org/10.1895/wormbook.1.4.1

Priess, J. R., & Thomson, J. N. (1987). Cellular interactions in early C. elegans embryos. *Cell*, *48*(2), 241–250.

Robertson, S. M., Shetty, P., & Lin, R. (2004). Identification of lineage-specific zygotic transcripts in early Caenorhabditis elegans embryos. *Developmental Biology*, *276*(2), 493–507. http://doi.org/10.1016/j.ydbio.2004.09.015

Satija, R., Farrell, J. A., Gennert, D., Schier, A. F., & Regev, A. (2015). Spatial reconstruction of single-cell gene expression data. *Nature Biotechnology*, *33*(5), 495–502. http://doi.org/10.1038/nbt.3192

Sawyer, J. M., Glass, S., Li, T., Shemer, G., White, N. D., Starostina, N. G., et al. (2011). Overcoming redundancy: an RNAi enhancer screen for morphogenesis genes in Caenorhabditis elegans. *Genetics*, *188*(3), 549–564. http://doi.org/10.1534/genetics.111.129486

Schaner, C. E., & Kelly, W. G. (2006). Germline chromatin. *WormBook : the Online Review of C. Elegans Biology*, 1–14. http://doi.org/10.1895/wormbook.1.73.1

Schierenberg, E., & Strome, S. (1992). The establishment of embryonic axes and determination of cell fates in embryos of the nematode Caenorhabditis elegans. *Seminars in Developmental Biology*, *3*, 25–33. http://doi.org/10.1234/12345678

Seydoux, G., & Fire, A. (1994). Soma-germline asymmetry in the distributions of embryonic RNAs in Caenorhabditis elegans. *Development (Cambridge, England)*, *120*(10), 2823–2834.

Shaffer, S. M., Wu, M.-T., Levesque, M. J., & Raj, A. (2013). Turbo FISH: a method for rapid single molecule RNA FISH. *PloS One*, *8*(9), e75120. http://doi.org/10.1371/journal.pone.0075120

Sommermann, E. M., Strohmaier, K. R., Maduro, M. F., & Rothman, J. H. (2010). Endoderm development in Caenorhabditis elegans: the synergistic action of ELT-2 and -7 mediates the specification→differentiation transition. *Developmental Biology*, *347*(1), 154–166. http://doi.org/10.1016/j.ydbio.2010.08.020

Sönnichsen, B., Koski, L. B., Walsh, A., Marschall, P., Neumann, B., Brehm, M., et al. (2005). Full-genome RNAi profiling of early embryogenesis in Caenorhabditis elegans. *Nature*, *434*(7032), 462–469. http://doi.org/10.1038/nature03353

Subramaniam, K., & Seydoux, G. (1999). nos-1 and nos-2, two genes related to Drosophila nanos, regulate primordial germ cell development and survival in Caenorhabditis elegans. *Development (Cambridge, England)*, *126*(21), 4861–4871.

Sulston, J. E., Schierenberg, E., White, J. G., & Thomson, J. N. (1983). The embryonic cell lineage of the nematode Caenorhabditis elegans. *Developmental Biology*, *100*(1), 64–119.

Trapnell, C., Cacchiarelli, D., Grimsby, J., Pokharel, P., Li, S., Morse, M., et al. (2014). The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nature Biotechnology*, *32*(4), 381–386. http://doi.org/10.1038/nbt.2859

Walston, T., Tuskey, C., Edgar, L., Hawkins, N., Ellis, G., Bowerman, B., et al. (2004). Multiple Wnt signaling pathways converge to orient the mitotic spindle in early C. elegans embryos. *Developmental Cell*, *7*(6), 831–841. http://doi.org/10.1016/j.devcel.2004.10.008

Weismann, A. (1893). The germ-plasm. London: W. Scott.

Wieschaus, E. (1997). Nobel Lecture. In N. Ringertz (Ed.), *Nobel Lectures*. Singapore.

Woollard, A. (2005). Gene duplications and genetic redundancy in C. elegans. *WormBook : the Online Review of C. Elegans Biology*, 1–6. http://doi.org/10.1895/wormbook.1.2.1

Wormbase. WS170. 09 Feb 2007. http://ws170.wormbase.org/

Wormbase. WS252. 04 Dec 2015. http://ws252.wormbase.org/

WormMine. WS250. 29 Oct 2015. http://wormbase.org/tools/wormmine/

Xue, Z., Huang, K., Cai, C., Cai, L., Jiang, C.-Y., Feng, Y., et al. (2013). Genetic programs in human and mouse early embryos revealed by single-cell RNA sequencing. *Nature*, *500*(7464), 593–597. http://doi.org/10.1038/nature12364

Yamanaka, A., Yada, M., Imaki, H., Koga, M., Ohshima, Y., & Nakayama, K.-I. (2002). Multiple Skp1-related proteins in Caenorhabditis elegans: diverse patterns of interaction with Cullins and F-box proteins. *Current Biology : CB*, *12*(4), 267–275.

Yamanaka, S. (2007). Strategies and new developments in the generation of patient-specific pluripotent stem cells. *Cell Stem Cell*, *1*(1), 39–49. http://doi.org/10.1016/j.stem.2007.05.012

Yan, L., Yang, M., Guo, H., Yang, L., Wu, J., Li, R., et al. (2013). Single-cell RNA-Seq profiling of human preimplantation embryos and embryonic stem cells. *Nature Structural & Molecular Biology*, *20*(9), 1131–1139. http://doi.org/10.1038/nsmb.2660

Zeisel, A., Muñoz-Manchado, A. B., Codeluppi, S., Lönnerberg, P., La Manno, G., Juréus, A., et al. (2015). Brain structure. Cell types in the mouse cortex and hippocampus revealed by single-cell RNA-seq. *Science (New York, N.Y.)*, *347*(6226), 1138–1142. http://doi.org/10.1126/science.aaa1934

Zhu, J., Hill, R. J., Heid, P. J., Fukuyama, M., Sugimoto, A., Priess, J. R., & Rothman, J. H. (1997). end-1 encodes an apparent GATA factor that specifies the endoderm precursor in Caenorhabditis elegans embryos. *Genes & Development*, *11*(21), 2883–2896.

CHAPTER 3: PROFILING TRANSCRIPTS OF MULTIPLE DIVERGING GASTRULATING CELL TYPES IN THE *C. ELEGANS* EMBRYO TO IDENTIFY GENETIC REGULATORS OF MORPHOGENESIS

The work described in this chapter will make up the first half of a paper not yet completed. Ultimately the paper will describe a dissection and RNA-seq strategy to identify transcripts that correlate with morphogenesis in multiple cells types. Candidate genes will be tested in the embryo for their role in C. elegans gastrulation. This document describes the work through candidate gene discovery. Within the next year Mark Slabodnick, and perhaps other members of the Goldstein Lab, will perform follow up experiments with some of the candidate genes identified, and will complete the second half of the paper.

The concept and experiments for this paper were designed by me, with input from Bob Goldstein. All dissections, RNA-seq library preparation, and computational analysis were performed by me. The three transgenic strains in Figure 3.2 were generated by Mark Slabodnick. This half of the paper was written by me with feedback from Bob Goldstein.

Early in embryonic development, cells begin moving around each other to form the shapes of the body— a process called morphogenesis. Failures in morphogenesis can result in birth defects, and the inappropriate activation of morphogenesis programs can contribute to cancer metastasis (Wallingford et al., 2013, Blanco et al., 2002). It has been a long-standing challenge to identify the links between genetic patterning of the embryo and the mechanics of morphogenesis, possibly complicated by pleiotropic effects or partial redundancy with other pathways (Sawyer et al., 2011, Wieschaus 1997, Johnsen and Baillie, 1997, Gilmour et al., 2017). To search for transcripts involved in the initiation of morphogenesis, we performed RNA-seq on gastrulating cell types dissected from *C. elegans* embryos. *C. elegans* is a uniquely advantageous system for studying the genetics of morphogenesis, because the timing and position of every cell movement can be precisely anticipated (Sulston et al., 1983). This

stereotyped development allows us to isolate cells that are preparing to internalize— a state that is otherwise undetectable by eye. We compared transcriptomes of several different cell lineages over time that are each diverging in fate but similarly gastrulating (Harrell and Goldstein, 2011). We have identified 99 genes whose transcript abundance patterns reflect gastrulation across the various internalizing cell types. Interested in families of genes whose members may each be fulfilling similar roles in different cell types, we expanded our search to include groups of paralogous genes. We identified 51 groups of genes similar in sequence whose collective transcript abundance patterns correlate with gastrulation. We are now pursuing several candidates by genetic manipulation and microscopy, to test for potential roles in morphogenesis regulation.

3.1 INTRODUCTION

Morphogenesis, the process by which embryonic cells shape the body, requires an incredible coordination between cells (Trinkhaus, 1984). During phenomena such as neurulation, gastrulation, lung branching and heart looping, entire swaths of cells simultaneously move or change their shape, thereby reshaping tissues that can be orders of magnitude larger than any of the individual cells. Though each of these events takes place in different cell types, they use many of the same physical mechanisms. For example, both neurulation in the frog and gastrulation in the worm rely on apical constriction. Furthermore, some of the genes implicated in neural tube closure defects in mammals lead to gastrulation defects when disrupted in *C. elegans* (Sullivan-Brown et al., 2016). These morphogenetic events involve physical changes in the cell, such as assuming polarity, modulating adhesion, and locally activating the cytoskeleton (Purnell, 2013).

There has long been interest in understanding the links between genetic patterning of the embryo and the mechanics of morphogenesis (Wieschaus 1997, Sawyer et al., 2011, Gilmour et al., 2017). A small group of genes whose embryonic transcription results in a morphogenetic behavior are known, setting a precedent for the type of genes we might expect in *C. elegans*. One of these known genes is SNAIL, which has been shown to be responsible for epithelial-mesenchymal transition (EMT) in many animals including flies (Leptin, 1991), mice (Carver et al., 2001), chicken (delBarrio and Nieto, 2002) and urchin (Shu-Yu and McClay, 2007, Nieto, 2002). Another example is SHROOM3, an actin-binding protein that recruits ROCK to apical junctions (Hildebrand and Soriano, 1999, Nishimura and Tackeichi, 2008), and whose expression is sufficient to induce apical constriction in polarized epithelial cells (Haigo et al., 2003). We hypothesized that more genes like this— master regulatory genes for diverse morphogenetic events within an embryo and even across species— exist and could be uncovered with a strategic RNA-seq approach.

Morphogenetic events are critical for multicellular life, and an understanding of their genetic cues would provide critical insight into two causes of human disease, (1) when these genes are disrupted leading to birth defects (such as neural tube defects and congenital heart disease)(Wallingford et al., 2013, Srivvastava, 2006), and (2) when these genes are activated at an inappropriate time, contributing to the spread of cancer (such as the role of snail in EMT and tumor progression) (Blanco et al., 2002, Batlle et al., 2000).

C. elegans is a uniquely advantageous system for studying morphogenesis because of its invariant development. Work published by John Sulston and colleagues starting in 1983 allows researchers to predict exactly which cells will move, at what moment, and to where (Sulston et

al., 1983). A challenge that remains in the wake of Sulston et al., 1983 is to identify the molecular players that drive these morphogenetic movements.

One of the hypotheses for why regulators of morphogenesis have been difficult to identify is because many of these genes may have partially redundant functions with genes in other pathways (Wieschaus 1997). Traditional genetic screens are not well-suited to identify these categories of genes, because disruption of a single gene might not result in a noticeable phenotype (Johnsen and Baillie, 1997). Genomic techniques such as RNA-seq can help fill in these blind spots by taking a snapshot of the transcript level of every gene, regardless of whether they are partially redundant with other pathways or not. However, the asynchronous development of *C. elegans* embryos has presented a challenge to those wishing to perform RNA-seq on purified cell types or embryonic stages. Without a tractable method for synchronizing embryos, bulk embryonic tissue samples are most often sourced from mixed stages. The recent availability of reliable low-input RNA-seq protocols (Svensson et al., 2017, Adiconis et al., 2013) provides a solution to the challenge of purifying cell types from mixed-stage embryos: A sample collected and sequenced from a single embryo can be precisely identified in time and space.

The *C. elegans* embryo, with its stereotyped development, provides a uniquely powerful venue for studying the initiation of morphogenesis. At the 26-cell stage in the *C. elegans* embryo, the E lineage (endoderm precursor cells) moves from the outside of the embryo to the inside, marking the beginning of gastrulation (Nance et al., 2005). When transcription is inhibited in the *C. elegans* embryo, the E cells do not internalize (Powell-Coffman et al., 1996). Furthermore, when *end-1* and *end-3* transcripts (transcription factors that establish endodermal

fate) are disrupted via RNAi, the E cells fail to internalize (Maduro et al., 2005). This suggests that gastrulation is transcriptionally regulated in the worm, and that it is possible to identify regulators of morphogenesis by investigating changes in the transcriptomes of gastrulating *C. elegans* cells at the time points leading up to gastrulation.

After the two E cells internalize, several other cell types follow. Studying the transcriptome of only the E lineage would lead to ambiguous results, as transcripts that regulate morphogenesis would be indistinguishable from transcripts that regulate endoderm fate. Rather than studying just one internalizing cell type, we decided to compare transcriptomes across several internalizing cell lineages as they adopt different fates but initiate similar morphogenetic processes. We hypothesized that we would find transcripts that become enriched over time in cell types that are preparing to internalize, but not in cell types that do not internalize. Furthermore, we hypothesized that disrupting those transcripts would compromise the regulation of this morphogenetic event.

3.2 RESULTS

Dissection and Low-Input RNA-seq of Diverse Internalizing Cells in the Early *C. elegans* Embryo

We selected 4 groups of internalizing cells to study— MS descendants, E descendants, D descendants and descendants of Cap and Cpp, hereafter referred to as Cxp descendants (Figure 3.1A). We also selected two non-internalizing groups as negative controls— ABp descendants (of which only 2 out of 32 internalize) and Cxa descendants (none of which internalize) (Harrell and Goldstein, 2011). Cxa descendants would be the most challenging samples to dissect and prepare for RNA-seq because of their small size and number, but they would provide a powerful

dataset when compared to Cxp, internalizing cells with a close lineal relationship but a different cell behavior.

For each cell type, we were interested in identifying transcripts whose abundance increased during the cell cycles leading up to internalization. To do this, we needed two transcriptomes from each internalizing cell lineage— one before internalization, and one during. For the "before" time point we chose 2-3 cell cycles before cell internalization. To exclude transcripts that become broadly enriched in all cell types over the course of development, regardless of morphogenetic behavior, we sampled the two non-internalizing negative control cell types, choosing time that matched the timing of most of our other samples (Figure 3.1A).

In a prior study, we had already generated transcriptomes for 6 of the 11 samples of interest (Tintori et al., 2016). The five remaining samples without previously generated transcriptomes were those from around the 100 cell stage. From here on we will refer to this later stage as the 100-cell stage, even though the timing of the samples collected vary a bit, based on the timing of cell internalization (ABp and MS: 96 cell-stage, D: 108 cell stage, C: 212 cell stage, Figure 1A). We generated strains with promoter fusion fluorescent markers to mark each cell type of interest, and dissected single embryos by hand on a spinning disc confocal microscope. This technique allowed us to identify cell types by low levels of fluorescence, and collect them immediately after isolation from the embryo. We used a P*ten-1*::GFP transgenic strain to identify and collect ABp descendant cells ("ABp+") and the non-ABp remainder of each embryo ("ABp-") from matching embryos (Drabikowski et al., 2005). We generated a P*ceh-51*::mNG transgenic strain to sample MS descendants ("MS+") and the non-MS remainder of each embryo ("MS-"), and a P*fkh-2*::mNG transgenic strain to collect D descendants ("D+") and the non-D remainder of each embryo ("D-"). We generated a PR02D3.1::mKate2 (marking C

descendants) x P*hlh-1*::mNG (marking MS and Cxp descendants) transgenic strain to collect internalizing C descendants ("Cxp", or "C-in"), non-internalizing C descendants ("Cxa", or "C-out"), and the non-C remainder of each embryo ("C-"). We selected R02D3.1 as a marker for the C lineage based on its robust C-specific expression at the 16-cell stage, according to our previous study (Tintori et al., 2016, tintori.bio.unc.edu). For each replicate, the cell type of interest was identified by fluorescence, collected in one tube, and all or most of the remaining cells from the embryo were collected separately (Figure 3.1B). For each of the strains that we generated, we chose to generate promoter fusions and insert them at single copy into MosSCI sites in order to minimize any potential off-target effects of the transgenes (Frøkjaer-Jensen et al., 2008).

The transcriptomes already available from our previous study were generated from single cells (Tintori et al., 2016). Paralelling the methods in Tintori et al., 2016, we performed RNA-seq on these later time points using material from single embryos rather than bulk tissue. For the same reason, we used the kit and protocol that had been used for earlier samples. With the addition of these new datasets, we have now sequenced transcriptomes for all cell types of the 1-, 2-, 4-, and 8-cell stage, 9 cell types of the 16-cell stage (ABalx, ABarx, ABplx, ABprx, MSx, Ex, Cx, D and P_4), and 6 groups of cells from the 100 cell stage embryo (ABp+, MS+, D+, C-in, C-out, and an approximation of the remainder of the 100 cell stage embryo, referred to here as ABa-E-P_4) (Figure 3.1C). The ABa-E-P_4 transcriptome was generated in silico based on weighted averages of other samples from that stage (details in Experimental Methods). A previous study generated transcriptomes for many of these cell types by isolating founder cells starting at the 2- cell stage, and allowing cells to divide in culture before collecting them for RNA-seq (Hashimshony et al., 2015). The method used in that study kept cells naïve to critical

cell-signaling events. We were interested in preserving those fate-determining signaling events, so we collected cells based on fluorescence within 10 minutes of being disrupted from their native environment in the embryo.

Transcriptomes of Internalizing Cell Types Allow the Identification of Gastrulation-Related Transcript Abundance Patterns

Among 57 replicates of 9 samples from the 100 cell stage, we detected transcripts from 7,998 genes (above a threshold of 25 reads per kilobase of transcript per million mapped reads, or RPKM). This value roughly matched our expectations based on transcriptomes generated from earlier cell types— in our previous study we detected 8,575 genes amongst 1- to 16-cell stage embryos. In the previous study we thoroughly validated our low-input RNA-seq data by (1) comparing them to previously known gene expression patterns and (2) comparing them to single molecule fluorescent *in situ* hybridization assays (Tintori et al., 2016). The sequencing technology used in this study was virtually identical, with the main difference being that the samples collected were from groups of smaller cells later in development, rather than single larger cells earlier in development. To validate the dataset in the present study, we compared our mRNA sequencing data to previously reported protein level data specifically in the Cxa (C-out) and Cxp (C-in) samples. We chose these samples because they were the most technically difficult dissections, due to the cells' small size and low fluorescence levels, and hence were the samples we had the least confidence in.

Using the EPIC (Expression Patterns in Caenorhabditis) lineagomics database from Bob Waterston's lab (Sarov et al., 2012), we selected 6 genes, 3 of which (*cwn-1*, *tbx-8* and *vab-7*) showed protein expression in all C descendants, and 3 of which showed proteins expression in either Cxa descendants, the non-internalizing C lineage cells (*elt-1* and *nhr-171*) or Cxp

descendants, the internalizing C lineage cells (*egl-5*). *tbx-8* trascripts were detected in our dataset in both Cxa and Cxp descendents, as expected, but *cwn-1* and *vab-7* transcripts were only in Cxp descendants (Figure 3.4A-C). *elt-1* transcripts were detected at a higher abundance in Cxa than Cxp, as expected (Figure 3.4D). *nhr-171* transcripts were detected at a higher abundance in Cxa cells than in Cxp cells, as expected (Figure 3.4E). *egl-5* transcripts were not detected in either Cxp or Cxa (expected to have very low transcript abundance, but slightly higher in Cxp cells)(Figure 3.4F). The few instances in which our data did not align with the Waterston lab's lineagomics data could be due to false negatives in our data, which are to be expected with RNA-seq from such a small amount of starting material. Overall, the sequencing data matched our expectations as set by the Waterston Lab's database, especially in light of the fact the Waterston lab's data are based on protein expression, whereas our data reflect transcript abundance, which we do not assume to have direct proportionality.

To identify transcripts that are enriched in internalizing cell types, we filtered these 7,998 transcripts for just those that became enriched at least two-fold over time in at least two of the four internalizing cell types, and did not become enriched by more than two-fold in the two negative control non-internalizing cell types. This analysis yielded 839 genes.

Of the four internalizing cell types sampled, all but the E lineage will generate some muscle cells (all D descendants, all Cxp descendants, and 17/52 MS descendants will become body muscle)(Sulston et al., 1983). In order to avoid genes that are exclusively related to muscle fate, we filtered the 839 transcripts for those that become enriched over time in the E lineage. This reduced the number of transcripts to 445. The ABp lineage (mostly non-internalizing) splits from all the internalizing cells we sampled (all P_1 descendants) at the two cell stage, leaving the possibility that comparisons between ABp and internalizing cell types would highlight

transcripts that are differentially enriched at the two-cell stage, rather than an internalization-correlated enrichment. In order to avoid this bias, we filtered the 445 genes for only genes that are enriched by two-fold in the internalizing C descendants (Cxp) when compared to the non-internalizing C descendants (Cxa). This reduced our filtered list to 150 genes. To avoid genes whose expression is notably higher in non-internalizing cell types at earlier stages than in internalizing cell types, we filtered these 150 genes for only those whose maximum transcript abundance in any non-internalizing cell types did not exceed the transcript abundance in the internalizing cell types that that gene is enriched in by more than two-fold. This left us with 99 genes. (Figure 3.2A).

Of these 99 genes, 54 transcripts were enriched, as defined by the filtering above, in two of the four internalizing cell types, 33 were enriched in three of the four, and 11 were enriched in all four (as in the pattern shown in Figure 3.2B). These 11 transcripts were *cav-1* (encoding a caveolin homolog), *pcp-5* (a peptidase), *kbp-2* (kinetochore binding protein), *cya-1* (a cyclin), *rps-21* (a small ribosomal subunit), *fbxc-39* (an F-box C protein), *upb-1* (a hydrolase), *ver-3* (a VEGF receptor), *fbxb-78* (an F-box b gene), Y75B8A.14 (a GTPase with a human ortholog) and *klp-19* (a microtubule motor protein)(Wormbase, 2017). Transcript abundance patterns for these 11 genes are shown in Figure 3.2C.

Paralogous Transcripts Encoding LIM Domains Each Become Enriched in Different Internalizing Cell Types

We originally hypothesized that we could identify regulators of morphogenesis by looking for transcripts that are enriched in only internalizing cell types. Aware of the possibility that different members of a gene family may be fulfilling this role in different cell types, we expanded our analysis to include groups of genes are similar to each other in sequence.

We created groups of genes based on a protein BLAST E score cutoff of e-15 (see Experimental Procedures), and calculated a hypothetical transcript profile for each paralogy group by summing the transcript profiles of all the paralogous genes in the group (as shown in the last pictogram in Figure 3.3A). We evaluated each of these summed transcript profiles for correlation with gastrulation using the same steps we used to identify the 64 individual genes, above. We removed paralogy groups whose internalization-specific expression pattern could be attributed to a single gene that we had already identified. This analysis yielded 51 paralogy groups, 27 of which had transcripts that become enriched in two of the four internalizing cell types, 21 of which had transcripts that become enriched in three of the four internalizing cell types, and three paralogy groups whose transcripts become enriched by two-fold in all four internalizing cell types, when compared to their grandparent or great-grandparent cell type.

Of the three paralogy groups whose transcripts become enriched in all four internalizing cell types, two groups consisted of several genes that either had an F-box domain or no known domain at all. None of these genes have a known function or known orthologs outside of the Caenorhabditids. The third group consisted of three genes encoding proteins with LIM domains; *lim-9* (an ortholog of LIMPET in *Drosophila* and FHL2 in vertebrates), *pxl-1* (an ortholog of paxillin in *Drosophila* and vertebrates), and *zyx-1* (an ortholog of zyxin in *Drosophila* and vertebrates)(Wormbase, 2017). Within this paralogy group, *zyx-1* transcripts are enriched in the E lineage, *lim-9* transcripts are enriched in Cxp descendants, and *pxl-1* transcripts are enriched in MS descendants and D descendants. These genes are thought to encode proteins involved in actin filament organization, particularly in Z discs and dense bodies of muscle cells (Lecroisey et al., 2013, Kadrmas and Beckerle, 2004, Qadota et al., 2007, Warner et al., 2011). They have been implicated in focal adhesions, mechanotransduction (Cattaruzza et al., 2004, Hoffman et al.,

2012), stretch-induced gene expression (Nix and Beckerle, 1997), planar cell polarity and asymmetric cell division (Wu and Herman, 2006), and have been shown to interact physically with actin cytoskeletal components such as vinculin and alpha-actinin (Lecroisey et al., 2013). The enrichment patterns of these genes, their broad conservation across animals, as well as their known involvement in cytoskeletal organization, makes them intriguing candidates for regulators of cell behavior during gastrulation.

3.3 DISCUSSION

A Genomic Method for Identifying Potential Genetic Regulators of Morphogenesis

Identifying the genetic links between embryonic patterning by transcription factors and the cytoskeletal mechanisms of morphogenesis has been an on-going challenge (Wieschaus, 1997, Gilmour et al., 2017). One hypothesis is that there may be multiple genetic pathways whose functions overlap, making morphogenesis robust to the loss of just one pathway (Sawyer et al., 2011). Worms have been estimated to have a per-gene duplication rate 2.5x higher than yeast, and 10x higher than Drosophila, giving them an increased likelihood of genetic redundancy compared to other model organisms (Lynch and Conery, 2000, Woolard, 2005). This would create a scenario in which mutagenesis of a single gene involved in morphogenesis may not be detected as such, because morphogenesis would still be completed, even if it is compromised. In this study we have identified potential regulators of morphogenesis by scouring the transcriptomes of gastrulating and non-gastrulating cell types for genes whose expression coincides with gastrulation, regardless of potential redundancy.

Because gastrulation in the worm relies on embryonic transcription, we hypothesized that regulators of *C. elegans* gastrulation can be identified by changes in the transcriptome.

Morphogenesis may not necessarily be transcriptionally regulated in all animals, though. Morphogenesis may be regulated by maternal determinants, signaling, or tension sensors in the cells. Even so, we speculate that transcription-dependent regulators in the *C. elegans* embryo might have homologous mechanisms in other animals that serve similar functions but are regulated independent of transcription.

RNAi Targeting *zyx-1* Compromises Gastrulation in the E lineage

The three LIM domain containing proteins described above are broadly conserved, and are best known for actin and myosin organization in thick filament bundling (Hoffman et al. 2006, Kadrmas and Beckerle, 2004). Their proteins have been identified in M-lines, dense bodies and Z-discs in muscles, and sites of actin attachment in focal adhesions (Lecroisey et al., 2013, Qadota et al., 2007, Warner et al., 2011). We had considered that these transcripts might be muscle-specific, but two observations lead us to believe that they may be serving morphogenesis-specific functions at this point in embryogenesis. First, *zyx-1* transcripts are enriched in the E cell, whose descendants will only become intestine (Tintori et al., 2016, tintori.bio.unc.edu). Second, these data represent a very early stage of embryogenesis, with transcripts of these three genes detected between 90 and 210 minutes post-first cytokinesis. It is unlikely that muscles are beginning to function this early in embryogenesis. For comparison, *myo-3* and *unc-54*, body wall muscle myosins A and B, are first detected as transcripts at very low levels at 240 minutes, gradually increasing until reaching their peak transcript abundance at 690 minutes (Celniker et al., 2009).

Preliminary studies of *zyx-1*, performed by Goldstein lab masters student Tim Cupp, further support the hypothesis that these LIM containing proteins may be involved in

gastrulation. When Tim targeted *zyx-1* by dsRNA injection, embryos showed 17.6%, 28% and 11.5% embryonic lethality across three replicates with sample sizes of 34, 25, and 78 embryos, respectively (Cupp, 2017). By comparison, mothers injected with mock dsRNA showed embryonic lethality of 1.3%, 1.4%, and 0% (n=147,72,102). In typical embryos, E descendants internalize before dividing from two to four cells. When Tim targeted *zyx-1* transcripts by dsRNA injection, E descendants in 25% of embryos did not internalize before dividing into four cells (n=20).

During apical constriction of the internalizing E descendants in typical embryos, robust myosin movement begins before the apical surface begins to shrink (Roh-Johnson et al., 2012). This creates two phases of apical constriction— before the membrane becomes linked to the contracting myosin, and after (Figure 5.1A). In embryos in which Tim targeted *zyx-1* by dsRNA, the second phase of apical constriction showed much higher slippage between myosin and membrane than in wild type, suggesting that the contracting myosin was not effectively linking to the cell membrane (Figure 5.1B)(Cupp, 2017). Collectively, these data suggest that in E descendants, ZYX-1 may be important for linking contracting myosin to the cell membrane.

Future Directions – Testing Transcripts with Internalization-Specific Enrichment Patterns for Roles in Gastrulation

Starting with the three genes that encode LIM domains, identified above, we will begin testing candidates for potential roles in gastrulation. To test if these genes are necessary for internalization, we will disrupt gene expression via RNAi, and observe the embryos for cell internalization defects. The fluorescent marker strains shown in Figure 3.1B will facilitate this screen, because many of the fluorescent cells should be internalized by the end of gastrulation. Using these strains, we will inject dsRNA, incubate the embryos through gastrulation, and then

image the embryos to evaluate whether the fluorescent cells have internalized or not. If one of a group of genes is found to be necessary for internalization, we will then overexpress these genes in non-internalizing cells, for example the Cxa descendants, in order to test for sufficiency of these transcripts to induce ectopic internalization.

3.4 MATERIALS AND METHODS

Worm Husbandry and Embryo Dissections

All worms were grown at 20°C and dissected at room temperature (21°C– 24°C). Single embryos were selected at 10–20 min before the desired stage and dissected based on Edgar and Goldstein, 2012. Dissections were conducted by aspiration on a Yokogawa spinning disk confocal microscope under brightfield illumination, and cell types were identified by fluorescent markers.

RNA Preparation, Sequencing, and Analysis

cDNA was generated using the SMARTer Ultra Low RNA Input for Illumina Sequencing Kit, and sequencing libraries were prepared using the Nextera XT kit, both according to manufacturers' instructions. Identical reads were collapsed before analysis as in Tintori et al., 2016. Sets of samples from a single embryo were rejected if one or more library had an over-representation of ERCC spike in reads (if the number of ERCC spike in transcripts were more than one tenth the number of worm transcripts)(Baker et al., 2005), suggesting a degradation of worm RNA during sample preparation. Transcripts were considered "detected" if their RPKM value was above a threshold of 25. All fold change calculations were done on adjusted RPKM

values— raw RPKM values with 25 added to them— to avoid enriching for small differences between samples with low RPKM values.

Calculation of ABa-E-P4 Sample from 100-Cell Stage

Approximated transcriptomes for whole embryos at the 100-cell stage were generated by weighing and summing RPKM values from individual embryos. For example, the average of all ABp+ transcriptomes was multiplied by (1/4), the average ABp- transcriptome (generated from all non-ABp cells of that same embryo) was multiplied by (3/4), and the two values were summed. Weights were determined by the fraction of the whole embryo that each sample represented (MS+ = 1/8, D+ = 1/16, C-in=1/16, C-out=1/16). The ABa-E-P_4 transcriptome was calculated by subtracting weighted values of each non- ABa-E-P_4 cell type (ABp+, MS+, D+, C-in and C-out) from the whole-embryo transcriptome. Finally, the remaining RPKM values for each gene were multiplied by (16/9), the inverse of the fraction of the embryo represented by ABa-E-P_4.

Defining Paralogy Groups

Genes were BLASTed against the *C. elegans* EST collection with an e-value threshold of 10^{-15}. This cutoff was chosen based on *end-1* and *end-3*, a known example of paralogous genes that overlap in function. e = 10^{-15} was the most conservative cutoff that resulted in *end-1* and *end-3* appearing in each other's list of BLAST hits.

FIGURES

Figure 3.1 A

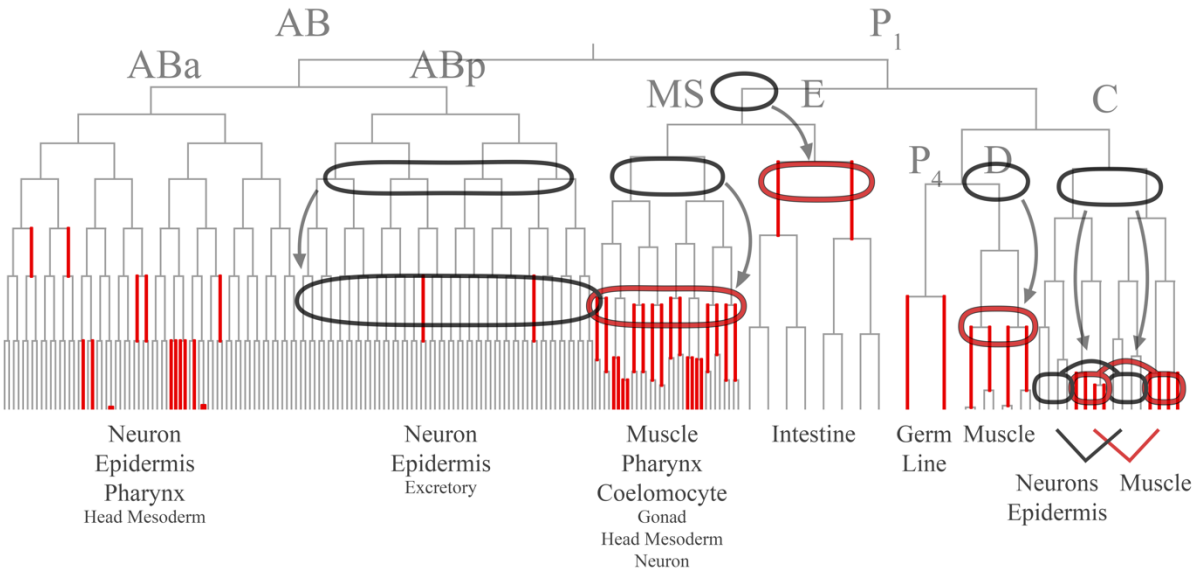

Figure 3.1B

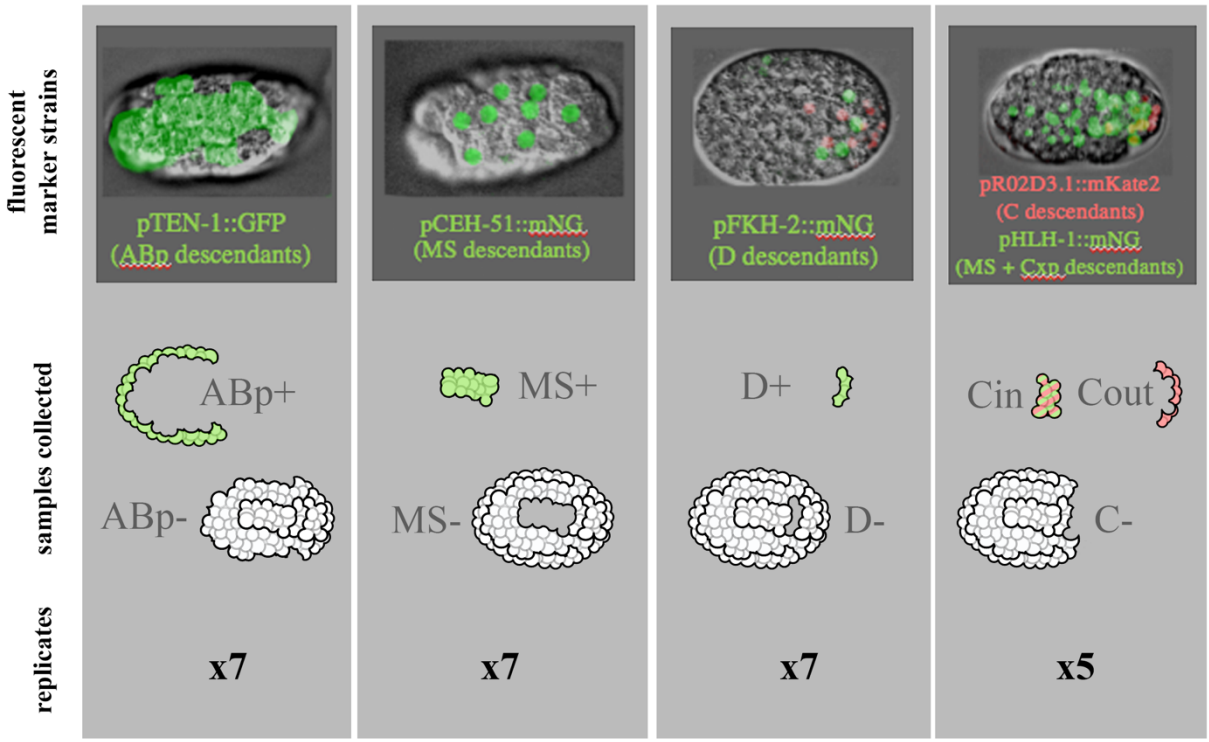

Figure 3.1C

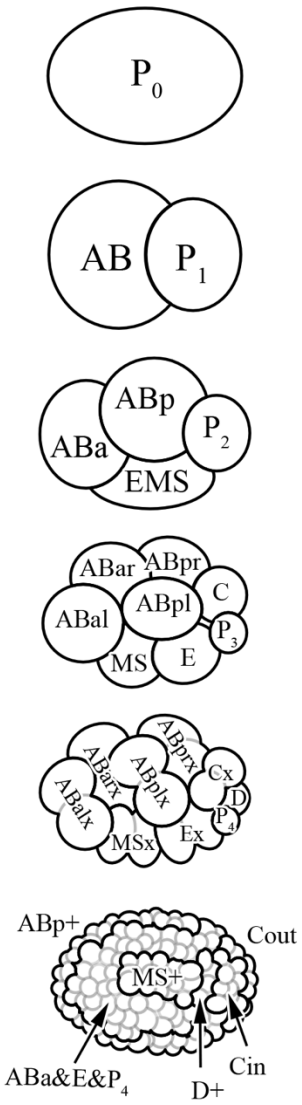

Figure 3.1 - Dissection and transcriptome profiling of multiple divergent, gastrulating cell types.

a. Samples were collected from 4 cell lineages that internalize during gastrulation— the MS, E, D and Cxp lineages (red branches), and 2 negative control lineages that do not internalize— the ABp and Cxa lineages (gray branches). For each lineage, a before and after sample was collected, with the "during" time point being at the start of internalization, and the "before" time point being 2-3 cell cycles earlier.

b. Four fluorescent marker strains that were used to dissect and collect each sample collected from the 100 cell stage.

c. Key of cell types for which transcriptomes were generated in Tintori et al. 2016 and in the present study.

Figure 3.2A

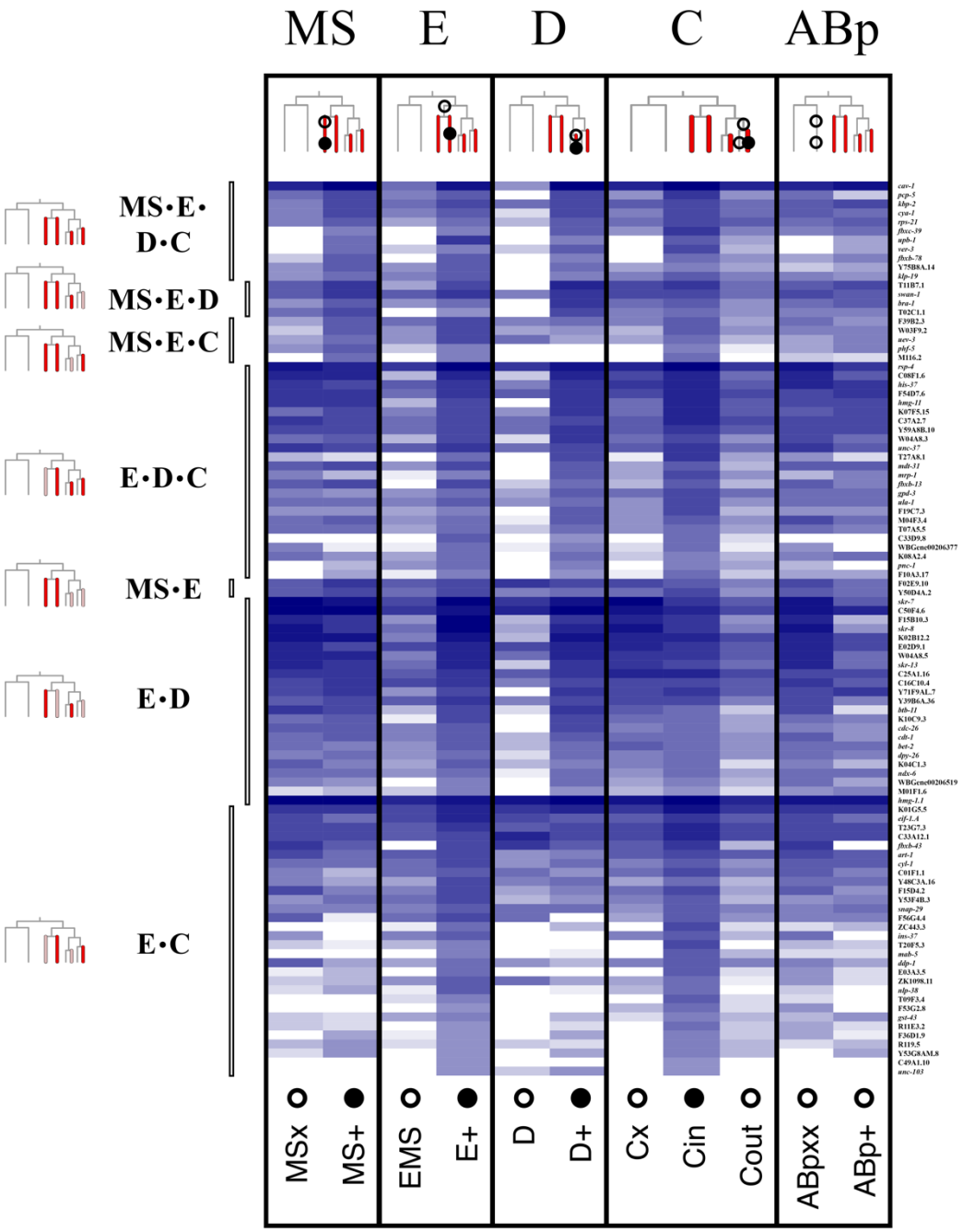
MS
E
D
C
ABp
MS·E·D·C
MS·E·D
MS·E·C
E·D·C
MS·E
E·D
E·C
MSx
MS+
EMS
E+
D
D+
Cx
Cin
Cout
ABpxx
ABp+

Figure 3.2B&C

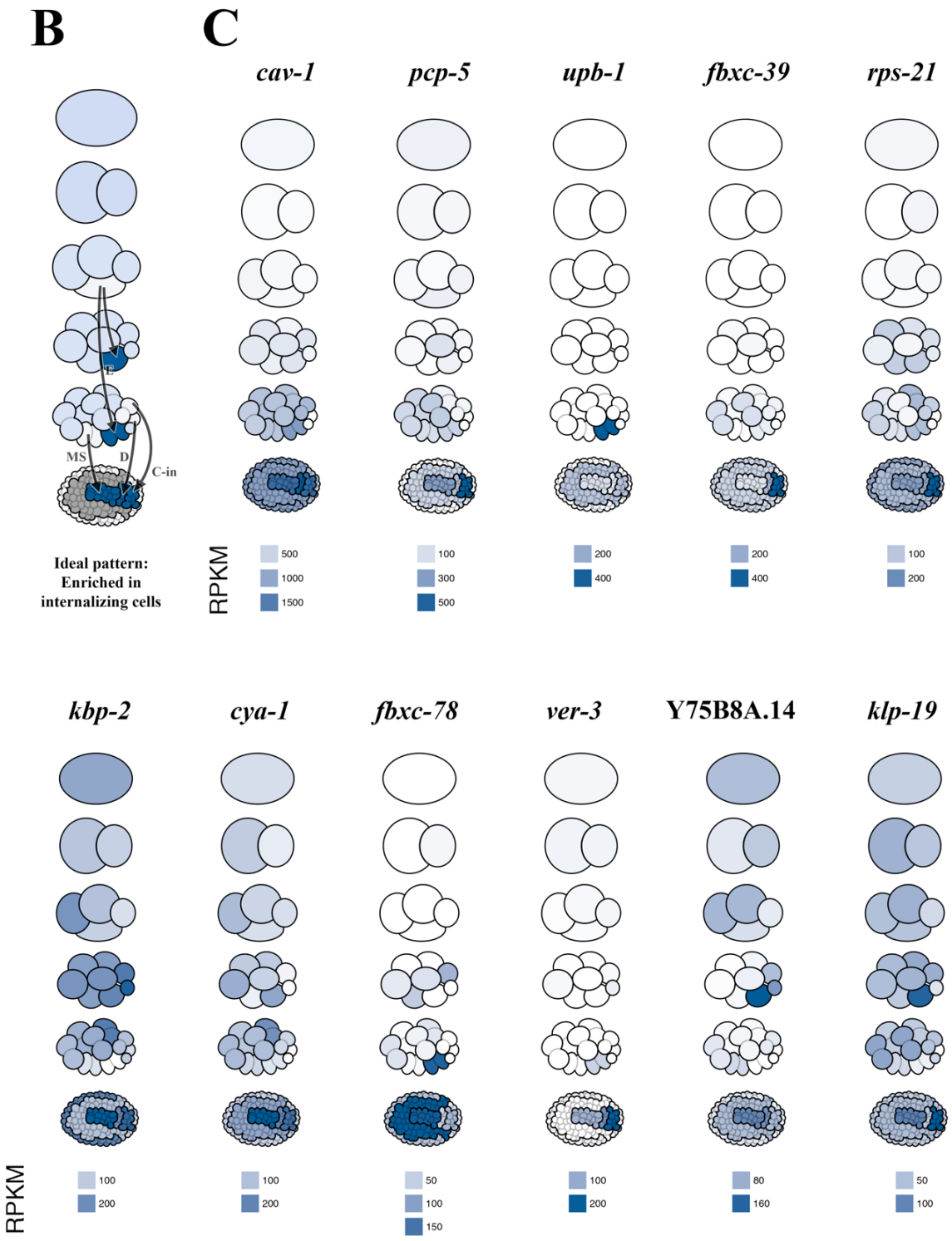
B
C
cav-1
pcp-5
upb-1
fbxc-39
rps-21
E
MS
D
C-in
Ideal pattern:
Enriched in
internalizing cells
RPKM
500
1000
1500
100
300
500
200
400
200
400
100
200
kbp-2
cya-1
fbxc-78
ver-3
Y75B8A.14
klp-19
RPKM
100
200
100
200
50
100
150
100
200
80
160
50
100

Figure 3.2 - Low-input RNA-seq of early embryo reveals genes with gastrulation-enriched transcript abundance patterns.

a. Heatmap of transcript abundance of 99 genes in 11 samples from the 100 cell stage embryo. The transcripts shown are those whose abundances increase over time in the E lineage and at least one of the three other internalizing cell types, but not in the negative control non-internalizing cell types. These genes also have at least two-fold abundance in internalizing C cells when compared to non-internalizing C cells. The lowest transcript abundance in internalizing cells is no less than one half the highest transcript abundance in any non-internalizing cell. Each gene is categorized by how many internalizing lineages its transcripts are enriched in. At the top of the heat map, each internalizing lineage is diagrammed. Open circles represent samples of cells that are not internalizing, and closed circles represent those that are.

b. An example of the transcript abundance pattern described the "MS E D C" group in Figure 3.2A. Hypothetical transcript abundance is heat mapped onto a pictogram. Abundance is high in internalizing cell types (MS+, E+, D+, C-in, in dark blue), low in non-internalizing negative control cell types and "before" stages (white), relatively low in other non internalizing cell types (light blue), and gray in all cell types not used in this analysis.

c. Transcript abundances heat-mapped on to pictograms of the embryo (key in Figure 3.1C). The eleven transcripts show increased abundance over time in the four internalizing cell types sampled, but not in the two non-internalizing negative control cell types.

Figure 3.3

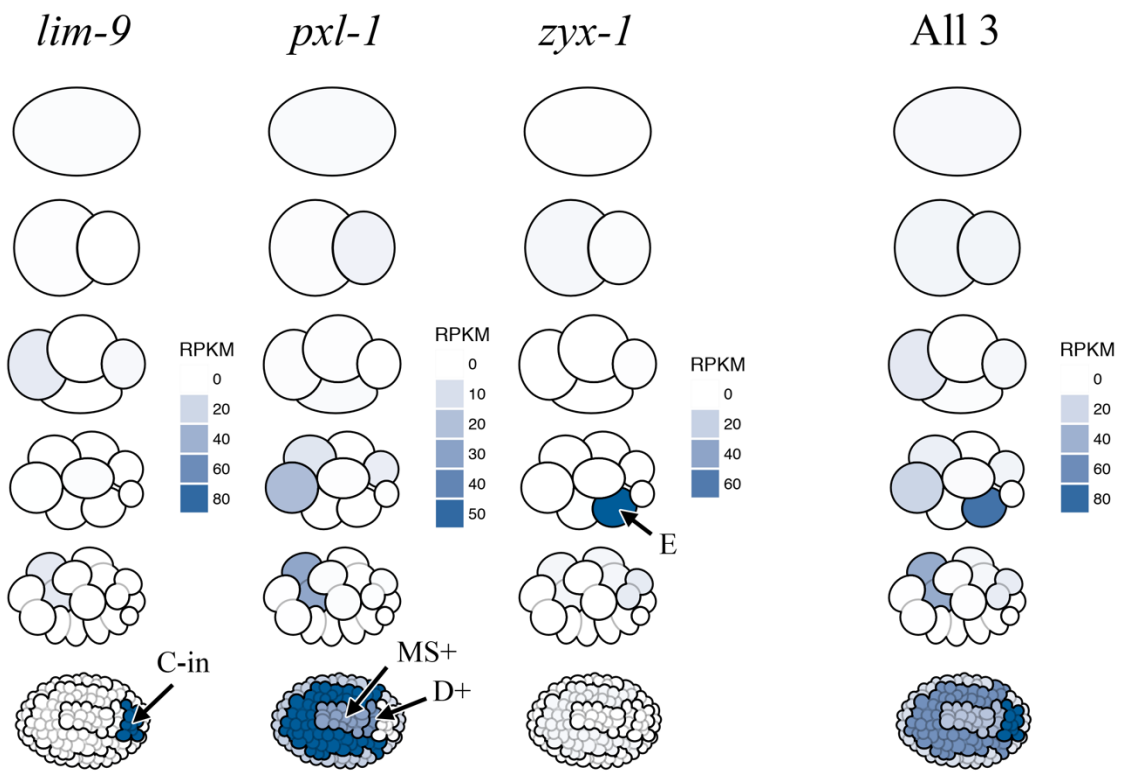

Figure 3.3 – Three paralogous LIM domain containing genes each show transcript enrichment in different gastrulating cell types.

a. Heat map pictograms of transcript abundance of three genes encoding LIM domains, and the expression pattern of the sum of their abundances. *lim-9* has increased expression over time in Cxp, the internalizing C descendants. *pxl-1* has increasing abundance in MS descendants and D descendants. *zyx-1* has increased abundance in E before internalization.

Figure 3.4: Supplement

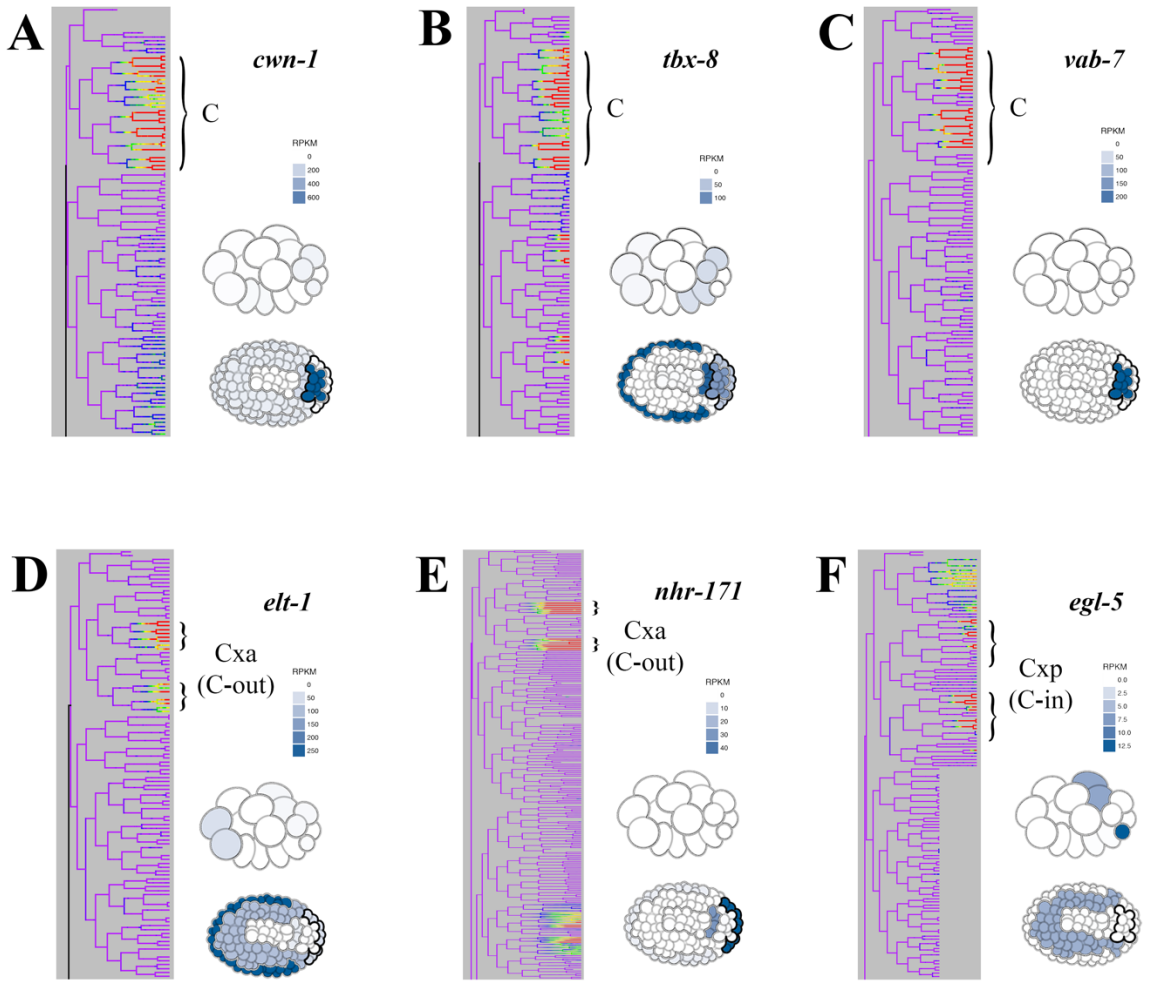

Figure 3.4: Supplement - Verification of C descendant transcriptomes.
Six genes with protein expression in C descendants were selected from the Waterston Lab's lineagomics database and compared to our transcriptiome data. Cell lineages show only the P_1 descendants, and are color coded by relative fluorescence levels detected from a film of embryonic development taken of embryos with multi-copy arrays of promoter fusions of each gene (Sarov et al., 2010). Transcripts of *cwn-1*, *fbx-8* and *vab-7* are expected to be enriched in all C descendants (outlined in black in A, B and C). Transcripts of *elt-1* and *nhr-171* are expected to be enriched in Cxa descendants, and transcripts of *egl-5* are expected to be enriched in decscendants of Cxp (outlines in black in D, E and F).

REFERENCES

Adiconis, X., Borges-Rivera, D., Satija, R., DeLuca, D.S., Busby, M.A., Berlin, A.M., Sivachenko, A., Thompson, D.A., Wysoker, A., Fennell, T., Gnirke, A., Pochet, N., Regev, A., & Levin, J.Z. (2013) Comparative analysis of RNA sequencing methods for degraded or low-input samples. Nature Methods 10(7):623–629.

Ahn D.G., Kourakis M.J., Rohde L.A., Silver L.M. & Ho R.K. (2002). T-box gene tbx5 is essential for formation of the pectoral limb bud. Nature 417(6890):754-8.

Andachi, Y. (2004). Caenorhabditis elegans T-box genes tbx-9 and tbx-8 are required for formation of hypodermis and body-wall muscle in embryogenesis. Genes Cells 9(4):331-44.

Baker, S. C., Bauer, S. R., Beyer, R. P., Brenton, J. D., Bromley, B., Burrill, J., et al. (2005). The External RNA Controls Consortium: a progress report. Nature Methods, 2(10), 731–734. http://doi.org/10.1038/nmeth1005-731

Batlle, E., Sancho, E., Francí, C., Domínguez, D., Monfar, M., Baulida, J. & García De Herreros A. (2000). The transcription factor snail is a repressor of E-cadherin gene expression in epithelial tumour cells. Nat Cell Biol 2(2):84-9.

Blanco, M.J., Moreno-Bueno, G., Sarrio, D., Locascio, A., Cano, A., Palacios, J. & Nieto, M.A. (2002). Correlation of Snail expression with histological grade and lymph node status in breast carcinomas. Oncogene 21(20):3241-3246. http://doi.org/10.1038/sj.onc.1205416

Carver E.A., Jiang R., Lan Y., Oram K.F. & Gridley T. (2001) The mouse snail gene encodes a key regulator of the epithelial-mesenchymal transition. Mol Cell Biol. 21(23):8184-8.

Cattaruzza M, Lattrich C, Hecker M (2004). Focal adhesion protein zyxin is a mechanosensitive modulator of gene expression in vascular smooth muscle cells. Hypertension 43, 726–730.

Cebra-Thomas J.A., Bromer J., Gardner R., Lam G.K., Sheipe H., Gilbert S.F. (2003). T-box gene products are required for mesenchymal induction of epithelial branching in the embryonic mouse lung. Developmental Dynamics 226(1) 82–90.

Celniker, S.E., Dillon, L.A., Gerstein, M.B., Gunsalus, K.C., Henikoff, S., Karpen, G.H., Kellis, M., Lai, E.C., Lieb, J.D., MacAlpine, D.M., Micklem, G., Piano, F., Snyder, M., Stein, L., White, K.P. and Waterston, R.H. (2009). modENCODE Consortium. Unlocking the secrets of the genome. Nature. 459(7249):927-30. http://doi.org/10.1038/459927a.

Cupp, T. (2017). In Vivo Exmination of the Molecular Mechanics Underlying Apical Constriction's Initiation in C. elegans Gastrulation (master's thesis). Collection: UNC-Chapel Hill Electronic Theses and Dissertations.

del Barrio, M.G. and Nieto M.A. (2002). Overexpression of Snail family members highlights their ability to promote chick neural crest formation. Development 129(7):1583-93.

Drabikowski, K., Trzebiatowska, A. & Chiquet-Ehrismann, R. (2005). ten-1, an essential gene for germ cell development, epidermal morphogenesis, gonad migration, and neuronal pathfinding in Caenorhabditis elegans. Dev Biol. 282(1):27-38.

Edgar, L. G., Wolf, N., & Wood, W. B. (1994). Early transcription in Caenorhabditis elegans embryos. Development, 120(2), 443–451.

Frøkjaer-Jensen, C., Davis, M.W., Hopkins, C.E., Newman, B.J., Thummel, J.M., Olesen, S.P., Grunnet, M. & Jorgensen, E.M. (2008). Single-copy insertion of transgenes in Caenorhabditis elegans. Nat Genet. 40(11):1375-83. http://doi.org/10.1038/ng.248.

Garrity, D.M., Childs, S. & Fishman, M.C. (2002). The heartstrings mutation in zebrafish causes heart/fin Tbx5 deficiency syndrome. Development 129(19):4635-45.

Gilmour, D., Rembold, M. & Leptin, M. (2017). From morphogen to morphogenesis and back. Nature. 541(7637):311-320. http://doi.org/10.1038/nature21348

Harrell, J.R. and Goldstein, B. (2011). Internalization of multiple cells during C. elegans gastrulation depends on common cytoskeletal mechanisms but different cell polarity and cell fate regulators. Dev Biol 350(1):1-12. http://doi.org/10.1016/j.ydbio.2010.09.012.

Hashimshony, T., Feder, M., Levin, M., Hall, B. K., & Yanai, I. (2015). Spatiotemporal transcriptomics reveals the evolutionary history of the endoderm germ layer. Nature, 519(7542), 219–222. http://doi.org/10.1038/nature13996

Hoffman, L.M., Jensen, C.C., Chaturvedi, A., Yoshigi, M. and Beckerle, M.C. (2012). Stretch-induced actin remodeling requires targeting of zyxin to stress fibers and recruitment of actin regulators. Mol Biol Cell 23, 1846–1859.

Hoffman, L.M., Jensen, C.C., Kloeker, S., Wang, A., Yoshigi, M. & Beckerle, M.C. (2006). Genetic ablation of zyxin causes Mena/VASP mislocalization, increased motility, and deficits in actin remodeling. J Cell Biol 172, 771–782.

Johnsen, R.C. and Baillie D.L. (1997). Mutation, in C. elegans II, edited by Riddle D. L., Blumenthal T., Meyer B. J., R. Priess J., editors. , pp. 79–95 Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY

Kadrmas, J.L. and Beckerle, M.C. (2004). The LIM domain: from the cytoskeleton to the nucleus. Nat Rev Mol Cell Biol 5, 920–931.

Lecroisey, C., Brouilly, N., Qadota, H., Mariol, M., Rochette, N. C., Martin, E., Benian, G. M., Segalat, L. S., Mounier, N. & Gieseler, K. (2013). ZYX-1, the unique zyxin protein of Caenorhabditis elegans, is involved in dystrophin-dependent muscle degeneration.. Mol Biol Cell, 24, 1232-49. http://doi.org/10.1091/mbc.E12-09-0679

Leptin M. (1991) twist and snail as positive and negative regulators during Drosophila mesoderm development. Genes Dev 5(9):1568-76.

Maduro, M.F., Hill, R.J., Heid, P.J., Newman-Smith, E.D., Zhu, J., Priess, J.R. and Rothman, J.H. (2005). Genetic redundancy in endoderm specification within the genus Caenorhabditis. Dev Biol. 284(2):509-22.

Lynch, M. and Conery, J.S. (2000). The evolutionary fate and consequences of duplicate genes. Science. 290:1151–1155

Nance, J., Lee, J.-Y. & Goldstein, B. (2005). Gastrulation in C. elegans. WormBook : the Online Review of C. Elegans Biology, 1–13. http://doi.org/10.1895/wormbook.1.23.1

Nieto MA. (2002). The snail superfamily of zinc-finger transcription factors. Nat Rev Mol Cell Biol 3:155-66.

Nix, D.A., Beckerle, M.C. (1997). Nuclear–cytoplasmic shuttling of the focal con- tact protein, zyxin: a potential mechanism for communication between sites of cell adhesion and the nucleus. J Cell Biol 138, 1139–1147.

Powell-Coffman, J.A., Knight, J. & Wood, W.B. (1996). Onset of C. elegans gastrulation is blocked by inhibition of embryonic transcription with an RNA polymerase antisense RNA. Dev Biol., 178(2):472-83.

Purnell B.A. (2013). Getting into Shape; Introduction to Special Issue. Science 340(6137) 1183. http://doi.org/10.1126/science.340.6137.1183

Qadota, H., Mercer, K. B., Miller, R. K., Kaibuchi, K. & Benian, G. M. (2007). Two LIM domain proteins and UNC-96 link UNC-97/pinch to myosin thick filaments in Caenorhabditis elegans muscle.. Mol Biol Cell, 18, 4317-26. http://doi.org/10.1091/mbc.E07-03-0278.

Roh-Johnson, M., Shemer, G., Higgins, C.D., McClellan, J.H., Werts, A.D., Tulu, U.S., Gao, L., Betzig, E., Kiehart, D.P. & B. Goldstein. (2012). Triggering a Cell Shape Change by Exploiting Pre-Existing Actomyosin Contractions. Science 335:1232-1235.

Sarov, M., Murray, J.I., Schanze, K., Pozniakovski, A., Niu, W., Angermann, K., Hasse, S., Rupprecht, M., Vinis, E., Tinney, M., Preston, E., Zinke, A., Enst, S., Teichgraber, T., Janette, J., Reis, K., Janosch, S., Schloissnig, S., Ejsmont, R.K., Slightam, C., Xu, X., Kim, S.K., Reinke, V., Stewart, A.F., Snyder, M., Waterston, R.H. & Hyman A.A. (2012). A genome-scale resource for in vivo tag-based protein function exploration in C. elegans. Cell 150(4):855-66. http://doi.org/1010.1016/j.cell.2012.08.001.

Sawyer, J. M., Glass, S., Li, T., Shemer, G., White, N. D., Starostina, N. G., Kipreos, E.T., Jones, C.D. & Goldstein, B.(2011). Overcoming redundancy: an RNAi enhancer screen for morphogenesis genes in Caenorhabditis elegans. Genetics, 188(3), 549–564. http://doi.org/10.1534/genetics.111.129486

Showell C., Binder O. & Conlon F.L. (2004). T-box genes in early embryogenesis. Dev Dyn. 229(1):201-18.

Shu-Yu W. and McClay D.R. (2007) The Snail repressor is required for PMC ingression in the sea urchin embryo. Development 134, 1061-1070. http://doi.org/10.1242/dev.02805

Srivastava, D. (2006). Making or Breaking the Heart: From Lineage Determination to Morphogenesis. Cell 126(6) 1037–1048. http://doi.org/10.1016/j.cell.2006.09.003

Sullivan-Brown, J.L., Tandon, P., Bird, K.E., Dickinson, D.J., Tintori, S.C., Heppert, J.K., Meserve, J.H., Trogden, K.P., Orlowski, S.K., Conlon, F.L. & Goldstein, B. (2016). Identifying Regulators of Morphogenesis Common to Vertebrate Neural Tube Closure and Caenorhabditis elegans Gastrulation. Genetics 202(1):123-39. http://doi.org/10.1534/genetics.115.183137.

Sulston, J. E., Schierenberg, E., White, J. G., & Thomson, J. N. (1983). The embryonic cell lineage of the nematode Caenorhabditis elegans. Developmental Biology, 100(1), 64–119.

Svensson, V., Vento-Tormo, R. and Teichmann, S.A. (2017). Exponential scaling of single-cell RNA-seq in the last decade. BioArXiv pre-print. arXiv:1704.01379v2.

Trinkaus, J.P. (1984). Cells into organs : the forces that shape the embryo. Prentice-Hall. Englewood Cliffs, N.J.

Wallingford, J.B., Niswander, L.A., Shaw, G.M. & Finnell, R.H. (2013). The Continuing Challenge of Understanding, Preventing, and Treating Neural Tube Defects. Science 339 (6123), 1222002. http://doi.org/10.1126/science.1222002

Warner, A.D., Qadota, H., Benian, G.M., Vogl, A. & Moerman, D. G. (2011). The Caenorhabditis elegans paxillin orthologue, PXL-1, is required for pharyngeal muscle contraction and for viability. Mol Biol Cell, 22, 2551-63. http://doi.org/10.1091/mbc.E10-12-0941

Wieschaus, E. (1997). Nobel Lecture. In N. Ringertz (Ed.), Nobel Lectures. Singapore.

Wilson V. and Conlon F.L. (2002). The T-box family. Genome Biol. 2002;3(6):REVIEWS3008.

Woollard, A. (2005). Gene duplications and genetic redundancy in C. elegans. WormBook, ed. The C. elegans Research Community, WormBook. http://doi.org/10.1895/wormbook.1.2.1.

WormBase web site, http://www.wormbase.org, release WS260, August 2017.

Wu, M., and Herman, M.A. (2006). A novel noncanonical Wnt pathway is involved in the regulation of the asymmetric B cell division in C. elegans.. Dev Biol, 293, 316-29. http://doi.org/10.1016/j.ydbio.2005.12.024.

CHAPTER 4 – SCIENCE STORIES TOLD WITH FILM

This chapter is a documentation of some of my past and current science communication projects, some ideas I have about the practice and philosophy of science communication, and plans for upcoming projects. Excerpts from this chapter may potentially be used in future grant applications or artist's statements.

4.1 A Personal History

I grew up in a community of storytellers. The writers and filmmakers that raised me taught me that there is nothing more influential than a story. They taught me that it is through hearing and telling stories that we connect with each other, are able to expand beyond our own personal experience, and decide what is important to us.

Biology is made of conflict and resolution, and these stories were what first drew me in. At first I was attracted to the animal stories, because those were the ones I could imagine a context for. But quickly I was absorbing narratives on every scale within the scope of biology. As I learned about molecular biology, I was compelled by the idea that every little enzyme, though it has no will, has many conflicting forces, be they ionic or steric, pushing and pulling on it, creating thousands of interwoven micro-dramas within the cell. I loved the pointillist arc of evolution, where hundreds or thousands of discrete changes can each be rationalized individually, but look unrecognizably complicated when one steps back. As I eventually became more familiar with what it actually takes to *do* science, I grew an appreciation for the stories of the scientists themselves— their struggles, accidental successes, and the psychology is takes to spend your life chasing huge and often unanswerable questions.

One of the most fulfilling parts of the decade I've spent in labs has been taking the stories that I have relatively exclusive access to, and sharing them with others. The topics that we study in biology are the topics that have fascinated people for as long as people have been capable of being fascinated— birth, death, sex, competition, communication, manipulation, growth, home, the body, what it means to be alive. Anyone can be moved by the drama, horror and comedy of the stories that make up biology, as long as those stories are told with the focus on conflict rather than vocabulary. If novels, paintings or philosophy offer us new ways of conceiving of these big ideas, so does biology. If a poem can make something snap in your heart, and open up a part of your brain you didn't know you had access to, so can biology. We have a constant hunger for new ways of thinking about the struggle of being alive. Biological stories are an often overlooked source, but they are happily received, if told well.

Before graduate school I began a number of storytelling projects to share animal or molecule stories with non-scientists. These projects were mostly in the form of short animated films, cartoon drawings, and short essays, and were distributed by podcast, blog, and a science column in a monthly newspaper (Figure 4.1). The journalistic rhythm of learning intensely about a topic, sharing what I learned with others, and then moving on to a new topic was very appealing to me. In 2012 when I joined Bob Goldstein's lab, he and I explicitly planned to consider my science communication projects part of my PhD. This unusual plan has allowed me to develop projects and ideas that are important to my unique career path. During the last year of my PhD., I have become involved in several new projects, now spanning a broader spectrum of audiences. In this chapter I will briefly discuss narrative structure and the challenges of applying

it to science stories, and then I will outline the three main audiences I am interested in making videos for in the near future.

4.2 Narrative Structure of the Scientific Story

In the classic narrative structure of the Hero's Journey, the main character is called to adventure and leaves their known realm (Act I), enters the unknown realm where they must face challenges, develop new skills, and reach some greater understanding of themselves (Act II), confront whatever it was that sent them on this journey, and finally return home, now a master of both worlds (Act III)(Figure 4.2)(Campbell, 1949). Myths, novels and narrative films generally follow this structure. Episodic TV follows this structure, too, but instead of emerging from the unknown realm a changed person, the hero must emerge unchanged, so that everything is back in place by the start of the next episode (Harmon, 2013). In my view, the principle misunderstanding about science is its narrative structure. The grand narrative loop of biology is almost entirely first act. One great question, one big "Wow, what?!" moment is huge enough to propel hundreds of researchers to work for decades or centuries. Of course, like a good movie, there are many little narrative loops within the overarching one. Each scene in a movie has a narrative structure of its own, just as each experiment conducted in a lab follows a narrative structure.

When we tell the story of science (in papers, in seminars, or in the news), we usually use the scientist as our main character, and focus entirely on the third act— the resolution to their question. For papers this is fine. Papers are supposed to be, amongst other things, an official documentation of all the boring details. For seminars, I believe that setting up a strong first act is the most important part, if you want people to have an emotional response to your words and

remember you. In journalism or other public-facing venues, focusing on the third act is a disaster. It is boring, alienating, and deeply misleads people to believing doing science is about **knowing things** rather than about **being confused**. The first act— the big "Whoa, what?!" moment— is what puts the audience and the scientist on the same page. The second act— when the scientist learns that they are shockingly unprepared to tackle the quest at hand, and must improve themselves before they can slay the dragon (cue "Rocky" training theme), is what gives the audience a relatable understanding of what the activity of doing science is.

Alternatively, the scientist doesn't have to be the hero at all. If the big question is about how a beetle finds its way home, let the beetle be the hero. But then you can't switch to plot points about what the scientist did. You must stick with the story you are telling, if you expect to retain the trust and attention of your audience.

4.3 Molecular Animation – Scientists Sharing Hypotheses with Colleagues

The processes we study in biology are dynamic in time and in space. Tissues twist and pinch and fold around each other to sculpt a body. Communities of animals and plants expand and shrink over mountain ranges. Chromatin makes a specific region of the genome accessible at one moment, but not at another. Most of these processes are too small, too large, or too obscured to see for ourselves. We rely primarily on words and static two-dimensional images to try to explain to ourselves and each other how we think these dynamic, invisible 3-D entities are behaving. This is an incredibly difficult task.

Even when one colleague successfully articulates their molecular hypothesis to another, it is likely that each colleague holds very different mental images of said hypothesis, based on unspoken details about the surrounding subcellular environment. There are many facts that we

know to be true about the inside of a cell (the high viscosity of water at the molecular scale, the high density of proteins in the cytoplasm, or size differences between molecular players) that are easy to forget, but may change our intuition about what we imagine is physically possible.

The power of animation is that it makes an invisible idea visible. With a well-crafted animation, one can get an impression of a concept, the characters involved, and be gently reminded of the parameters of the environment. This impression is immediate and it is intuitive. It does not have to be parsed from big words and disorienting sentence structures.

I have been teaching myself the basics of Maya, a 3D animation program (see a still from a very feeble first attempt in Figure 4.3), with the goal of making cinematic animations of scientists' molecular hypotheses. In October of 2017 I will be attending an advanced Maya course at a visual effects school in Los Angeles, where others will be learning how to animate for video games and Pixar movies. The specific topic of this course will be programming particles, or collections of objects to move randomly but within defined parameters (for example, animating smoke or a swarm of gnats). Maya is a program well-suited to molecular animation because one can program the (sometimes random) behavior of many objects with a single line of code, but it is more controllable than a simulation.

Computer animations can be alienating because of their synthetic feeling. One of my primary goals when learning Maya is to figure out how to animate molecules and subcellular environments to look like what they are— natural organic objects, each with their own differences and idiosyncrasies, rough edges and unique character. The perennial challenge with scientific animations is to balance the need for accuracy with the desire for a warm, inviting aesthetic.

Though these types of animation may be filed under 'science communication' at first glance, I believe that they will also serve a crucial role in the scientific process, particularly in hypothesis development. After all, it is only when one tries to draw something from memory that one realizes where the gaps in his or her knowledge are. Janet Iwasa, a molecular animator at University of Utah, has described collaborations in which her storyboards have led the researcher to ask important questions they wouldn't have thought to ask otherwise (Fleischman, 2009).

My hope is that by producing these animations, I will be able to (1) help scientists communicate and collaborate with each other, (2) help scientists fine-tune their hypotheses for themselves and (3) continue learning about new fields within biology by collaborating closely with those who are working to advance our understanding within those fields.

4.4 Narrative, Character Driven Science Communication – Scientists Sharing Their Work with Invested but Non-Scientific Experts

Many people are interested in hearing scientific stories, not only because of their inherent sense of curiosity, but for their work. Grantors must understand the work they are evaluating for funding. Policy-makers must understand the data that their policy ought to be based on.

This past winter I made a 5 minute video for a colleague's Beckman Young Investigators Award application (Figure 4.4). The challenge of producing a video for this specific audience was that the viewers were not experts on this topic, but they needed fully understand the research. This is different than a pop-science video where the primary goal is actually to inspire (Act I), and the details of the research matter almost not at all.

This type of audience is an attractive one for me, because it challenges me to be thorough and accurate but also inviting, accessible, and incredibly organized. To me, this is the magical position between being able to talk science and being able tell a story. This is not about making

art *inspired* by science, or making science that *looks artistic*. Its about fully understanding and being a fluent liaison between research and in visual storytelling. My ultimate interest in these types of narrative animations projects would be to make videos that help scientists and lawyers explain technical or scientific concepts to policy makers who need the information to make decisions about how the world should be run.

4.5 Documentary Film as Research

In addition to making technical animations to explain molecular hypotheses, and narrative videos to explain a concept or body of work, I am also currently co-directing a short documentary film. This film was commissioned by Smithsonian as part of a series about how southern Louisiana lives with water, and will premiere at the New Orleans Film Festival in October of 2017 (Figure 4.5).

As a scientist who had always lived in a community of artists, I'd heard noise for years about how similar the sciences and the arts are. I'd never been convinced. I always found the processes of doing each to be irreconcilably different. I'll concede that both practices strive to make sense of a confusing world. But the arts tend to focus on the human experience, while science goes to extraordinary lengths to exclude the human experience and its influence (for good reason). The arts describe subjective truths to make sense of the emotional world, whereas the sciences define objective truths about life's hardware.

But truth is a tricky idea, especially the objective kind. As scientists, we'll never know the boundless limits of how much we don't know. The genetic tools we use to study cells today were likely unimaginable 350 years ago when Antonie van Leeuwenhoek first saw a microorganism through a single lens microscope. What puzzles are we trying to solve today for

which we don't even have the tools to understand the full scope of the question? It is similar to trying to solve a jigsaw puzzle for which you don't know how many pieces are missing, but are pretty sure it's at least ten times what you have. The best we can do is collect the data we have access to, and try to figure out which pieces add up to a good story and which pieces can be left unresolved for now. With this in mind, it is very difficult to believe anyone when they argue that they know what is right. It is most probable that we are all wrong, in some way, about virtually everything.

Documentary film was the first discipline within the arts that I felt had a very similar process to research science. In documentary and in science, you cannot chose a topic that you already know the answer to. You can chose a topic that is fascinating and unresolved, and then poke at it for a long time to try understand as many sides of it as possible. In both disciplines, you must collect a huge amount of information, and then sit down to pick out the 3 to 7 data points that make the story you think is interesting and holds the most truth (be it literal or emotional truth, in the case of documentary). In both disciplines you can expect to walk away from a project with more questions than answers. And in both disciplines you must remain on alert for a story that is more interesting than the one you set out to tell, and be ready to switch directions to pursue that new story. While documentary has more flexibility to play with and comment on its own form, both research and documentary share the restriction of having to be based on truth, and the limitless complexities that come with truth.

4.6 Future Directions in Visual Storytelling

The three scales of projects I am pursuing and have described in this chapter— public-facing documentary, highly technical molecular animations, and short narrative explanatory

videos—run the spectrum of non-fiction visual storytelling that I am interested in. I am not yet sure where within this spectrum the next phase of my work will take me, but I believe that experience with each of these types of projects will inform the others, and make them stronger. Each project is designed to strengthen my ability to create strong bridges between people that are interested in, and even dependent on each other, but for whom a communication barrier exists.

4.7 Summary of Science Communication Projects Produced During Ph.D.

Stop Animation Videos

Six Tips for Achieving Invisibility: An interactive series of shorts stories about how to hide in the open ocean where there is nothing to hide behind.

Spiraling Fern Sperm: A short animation about how one cell can find another just by systematically changing the size of the spirals it swims in.

An unfinished series on the diversity of optical tricks used in vision across Metazoa.

Proprioception and Virtual Reality for Flies: A short animation for John Tuthill's Beckman Young Investigator Award application (University of Washington).

Illustration

Cytoskeleton GIFs: A series of hand-painted animated GIFs of various cytoskeletal phenomena, created for the Triangle Cytoskeleton Meeting.

Cover Art: Molecular Cell, Volume 51, Issue 6 September 26, 2013. Integral Nuclear Pore Proteins Bind to Pol III-Transcribed Genes and Are Required for Pol III Transcript Processing in *C. elegans* by Kohta Ikegami and Jason D. Lieb

Cover Art: Developmental Cell, Volume 38, Issue 4, August 22 2016. A Transcriptional Lineage of the Early *C. elegans* Embryo by Sophia Tintori, Erin Osborne Nishimura, Patrick Golden, Jason Lieb and Bob Goldstein.

CAHS hypothesis: Molecular animations of Thomas Boothby's hypothesis regarding the dynamics of Cytoplasmic Abundant Heat Soluble Proteins.

Education

Animation for Nanos: A weekend workshop at the New York Science Museum about cycles in animation and cycles in nature. I was invited to design and lead this workshop by Alexis Gambis, the creator of the Imagine Science Film Festival.

Tinto's Invertebrate of the Month: A science column in Mothers News, a monthly newspaper out of Providence, Rhode Island.

Eduard Buchner and the Death of Vitalism: A short historical animation about the study of fermentation, produced for Eric Landers Introduction to Biology MOOC out of MIT.

Documentary

Station 15: A short documentary that tells the story of the New Orleans drainage system through the perspective of high school senior and poet. This film was commissioned by Smithsonian/Louisiana Endowment for the Humanities. I am a co-director and animator for the film, which will premiere at the New Orleans Film Festival in October of 2017.

FIGURES

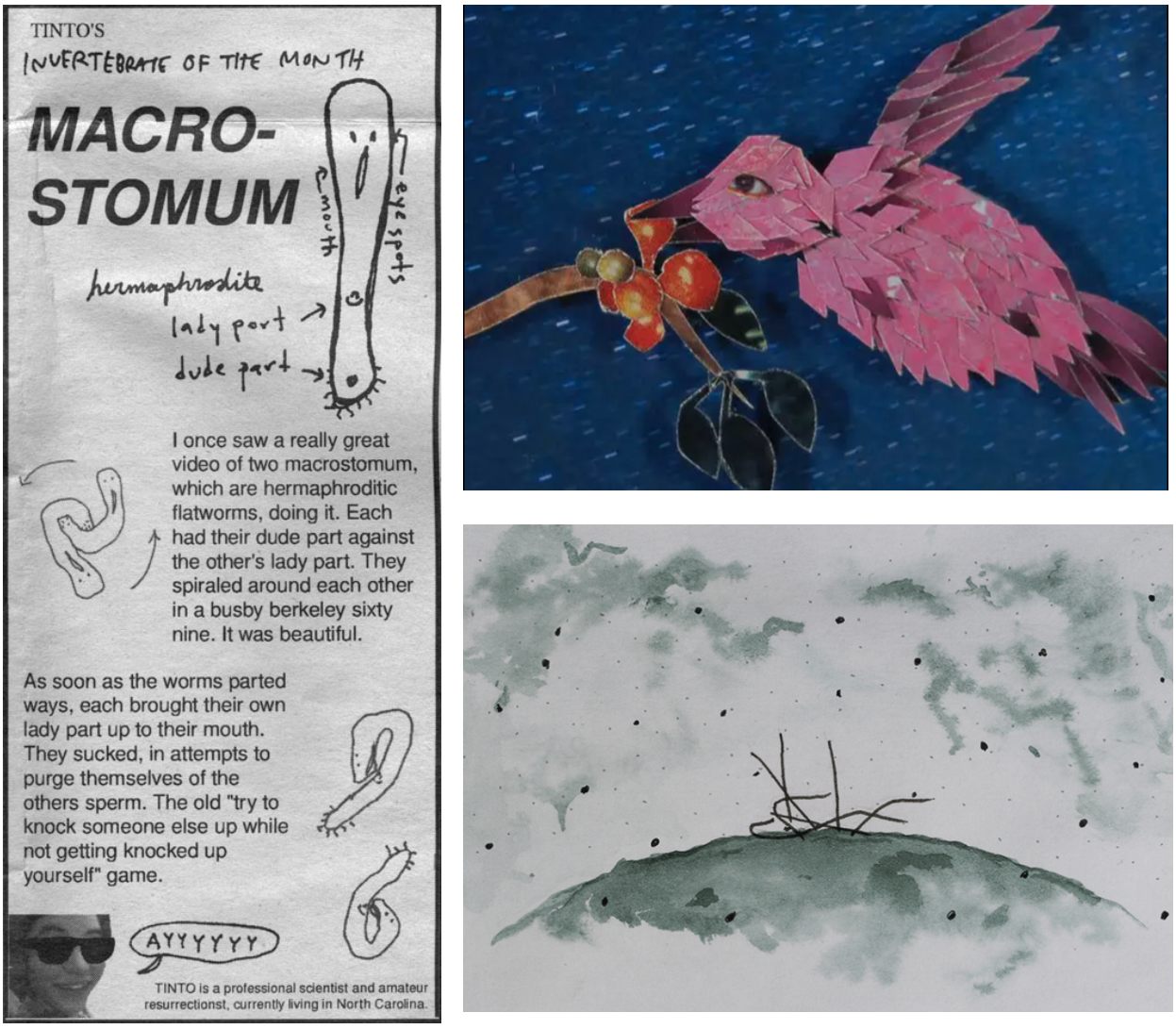

Figure 4.1 - Selection of images from previous science communication projects
Left – Except from a science column in Mothers News, a monthly newspaper out of Providence, RI. http://sophiatintori.com/invert.html
Top right – Still from a short animation about strangler figs from a series for CreatureCast.org http://sophiatintori.com/animations.html
Bottom right – Still from a series of animated gifs made for the Triangle Cytoskeleton Conference. http://sophiatintori.com/cytogifs.html

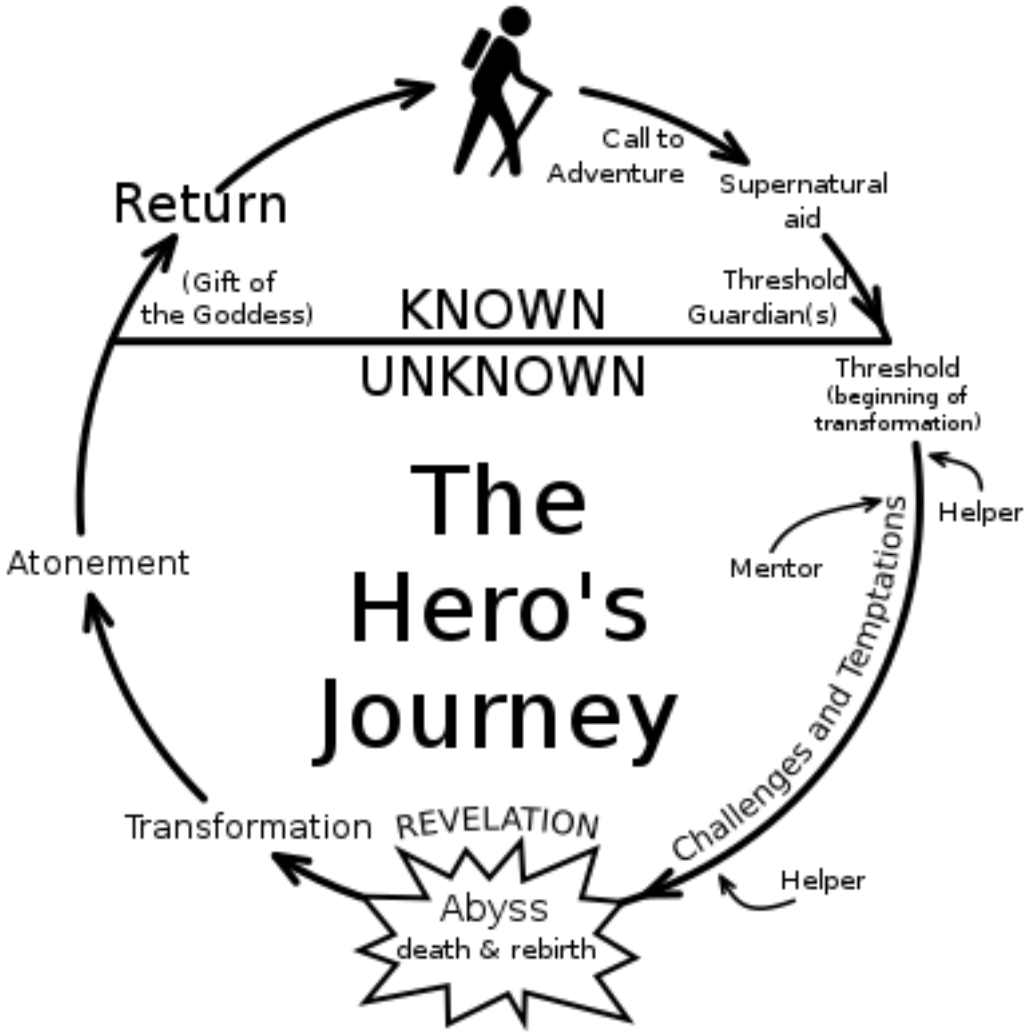

Figure 4.2 – Simplified diagram of the Hero's Journey. From Wikipedia Commons User :slashme, based on Campbell, 1949.

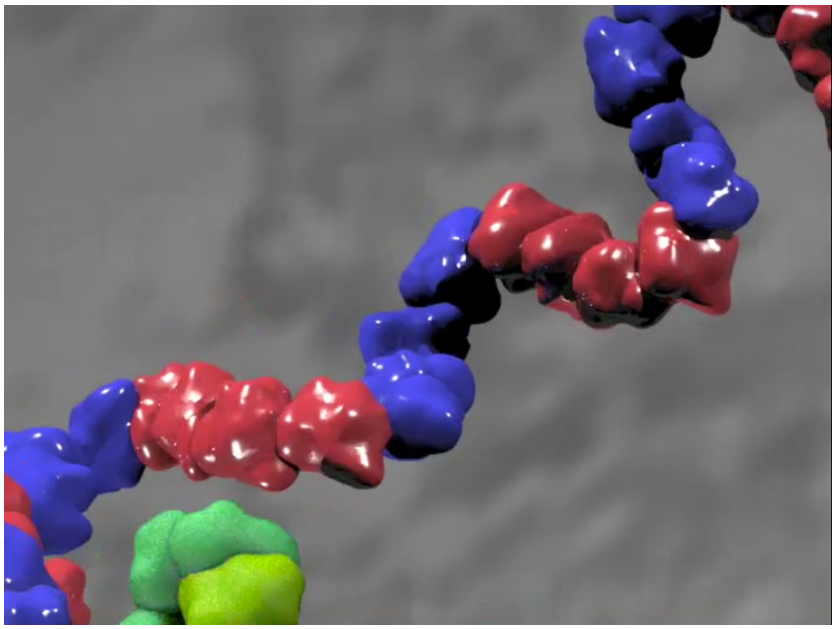

Figure 4.3 – Still from molecular animation of Cytoplasmic Abundant Heat Soluble proteins.
http://sophiatintori.com/molecularAnimation/CAHS.html

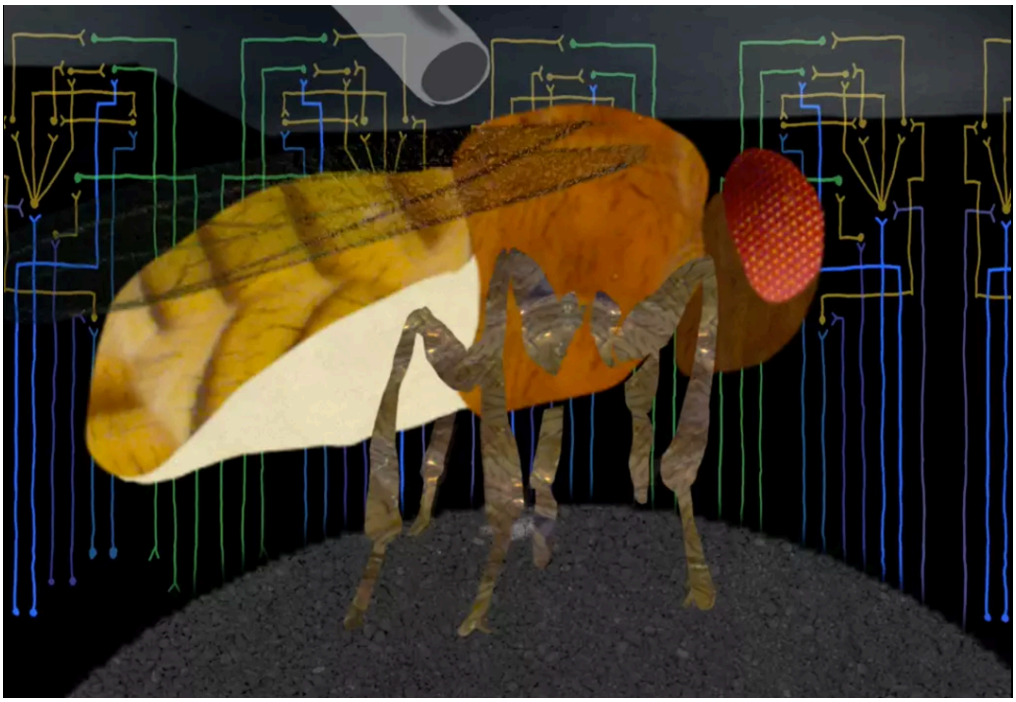

Figure 4.4 – Still from short narrative animation about proprioception.
http://sophiatintori.com/animations.html

Figure 4.5 – Four stills from short documentary film Station 15.

REFERENCES

Campbell, Joseph (1949). The Hero with a Thousand Faces. Princeton University Press. Princeton, N.J.

Fleichman, J. (2009). ASCB Profile: Janet Iwasa. ASCB Newsletter: 39–41.

Harmon, D. (2013). Story Structure 105: How TV is Different. Retrieved August 17, 2017, from http://channel101.wikia.com/wiki/Story_Structure_105:_How_TV_is_Different

CHAPTER 5: DISCUSSION

5.1 A Transcriptional Lineage of the Early *C. elegans* Embryo to Complement the Cell Lineage

For decades the *C. elegans* embryo has been a powerful tool for studying cell biology and development, largely because of its invariant cell lineage (Sulston et al., 1983). In Tintori et al., 2016, my colleagues and I described the transcriptomes of each cell of the embryo from the 1- to 16-cell stage, documenting the initiation of embryonic transcription. In the study described in Chapter 3, we generated transcriptomes for 11 more groups of cells, this time from the 100 cell stage. These data describe the genome-wide suite of transcripts present in early embryonic cells. Because all the cells sampled have a precisely known relationship to each other, these datasets allow a comparison of transcriptomes in space and time as these cells progressively diverge in fate, morphology, and behavior. As technology improves, scRNA-seq of cells beyond the 16-cell stage will become feasible, ideally allowing the possibility of a transcriptional map for every cell at every stage of development. The challenge of *post hoc* cell identification, explored in Hashimshony et al., 2012 and in Tintori et al. 2016, will continue to be relevant to these later stages of development. Until then, low input RNA-seq of *groups* of cells collected based on fluorescent markers, as in the 100-cell stage data described in Chapter 3, can be performed to complete a transcriptional lineage with a resolution slightly lower than single-cell.

5.2 Invariant Development of the Worm Embryo Allows Transcriptomes to be Identified *post hoc*

Several research groups have previously performed scRNA-seq on human and mouse cells, and identified their cell types *post hoc* by the transcriptomes (Grun et al., 2015; Yan et al., 2013; Biase et al., 2014; Xue et al., 2013; Zeisel et al., 2015; Jaitin et al., 2014; Trapnell et al., 2014; Satija et al., 2015; Achim et al., 2015; Pollen et al., 2014, Cao et al., 2017). The invariant development of *C. elegans* provides a constraint on the possible identities of each transcriptome. This advantage is not present in other systems and can help guide cell-type identification. For example, because each 4-cell embryo yields exactly one transcriptome each of exactly four cell types, we know that out of our total of 20 unidentified transcriptomes from this stage (4 cells x 5 replicates), exactly 5 of them are from P_2 cells, 5 are from EMS cells, 5 are from ABa cells, and 5 are from ABp cells. Because we know that a cell from one embryo will have exactly one counterpart in each other embryo, we were able to plot PCAs of all transcriptomes, as in Figures 2.2E and 2.2F, and look for clusters containing one cell from each embryo. Given the unique constraints of this system, such clustering suggests that all replicates of a single cell type are grouping together. The fact that known cell-type markers show transcript enrichment patterns that are consistent with our replicate grouping indicates the accuracy of our gene filtration and iterative PCA approach (Figures 2.2C, 2.2G, 2.2L, and 2.2Q).

Another difference between our study and previous scRNA-seq studies that identified cell types *post hoc* is that the *C. elegans* cells sampled in Tintori et al. 2016 divide about every 20 minutes, whereas the human and mouse cells of previous studies divide approximately every 24 hours. Given this comparatively short cell cycle, it is remarkable that the posterior *C. elegans* cells have such distinct transcript signatures, and by the same token perhaps not surprising that the anterior cells are difficult to distinguish from each other.

Previous studies have described transcriptomes at these stages of development at a lower spatial resolution (Baugh et al., 2003, 2005; Hashimshony et al., 2012; Hashimshony et al., 2015). By leaving embryos intact until immediately before sample collection, sequencing every individual cell at each stage, and using technology that captures full-length mRNAs, we have expanded upon these previous datasets. Our method preserves fate-determining cell-signaling events, allows for comparisons between groups of cells that were never sequenced separately before (such as AB descendants), and allows for inquiry into cell-specific variation in transcript splicing.

5.3 The Transcriptional Lineage Builds Upon Previous Studies of Early Embryonic Transcript Abundance

Previous studies that have measured transcript abundance in cells of the early embryo have either measured whole-embryo transcript levels (Baugh et al., 2003, 2005; Levin et al., 2012), or measured only parts of the embryo at a single-cell resolution and the rest of the embryo in clusters of related cells (Hashimshony et al., 2015). These clusters of cells were sampled by dissecting embryos starting at the 2-cell stage and allowing the isolated cells to divide in culture, then sequencing the group of descendants. This allowed descendants of founder cells to be harvested at later time points than in our study, but kept the cells naive to critical signaling events that take place in intact embryos. With the dataset presented in Chapter 2, by leaving all cells intact in the embryo until minutes before sampling, we captured single-cell transcriptomes while allowing the cell-cell signaling necessary for proper development to occur, and we detected the transcriptional results of this signaling (Figures 2.3B and 2.3F).

5.4 A Genomic Method for Identifying Regulators of Morphogenesis

Identifying the genetic links between embryonic patterning by transcription factors and the cytoskeletal mechanisms of morphogenesis has been an on-going challenge (Wieschaus, 1997, Gilmour et al., 2017). One hypothesis is that there may be multiple genetic pathways whose functions overlap, making morphogenesis robust to the loss of just one pathway (Sawyer et al., 2011). Worms have been estimated to have a per-gene duplication rate 2.5x higher than yeast, and 10x higher than Drosophila, giving them an increased likelihood of genetic redundancy compared to other model organisms (Lynch and Conery, 2000, Woolard, 2005). This could create a scenario in which mutagenesis of a single morphogenesis gene during a genetic screen may not be detected as such, because morphogenesis would still be completed, even if it is compromised. In the work described in Chapter 3, we have identified potential regulators of morphogenesis by scouring the transcriptomes of gastrulating and non-gastrulating cell types for genes whose expression is highly correlated with gastrulation, regardless of potential redundancy.

We hypothesize that regulators of *C. elegans* gastrulation can be identified by changes in the transcriptomes of gastrulating cells, because gastrulation in the worm relies on embryonic transcription. This direct connection between transcription and morphogenesis may not be present in many other organisms, though. Morphogenesis may be regulated by maternal determinants, signaling, tension sensors, or combinations of these. Even so, we speculate that transcription-dependent regulators in the *C. elegans* embryo might feasibly have homologous regulators in other systems that serve similar functions even if they may be regulated independent of transcription.

5.5 Transcript Profiles of Embryonic Cells Become Distinct From Each Other at Varying Rates

We have assigned a cell identity to each transcriptome from cells of the 1- to 16-cell stage embryo based on its transcript abundance data, cross-referenced to known expression patterns and in situ RNA hybridization. The transcriptomes of replicates of certain cell types (particularly the P_1 descendants) grouped together tightly, were clearly distinct from other cell types, and had identities confirmed by well-studied genes, making us confident in our assessments. For the anterior cell types whose transcriptomes were less distinct from one another, we have a lower confidence in our assignments (as in Figure 2.3D). We consider this paucity of distinguishing features to be an interesting biological result, suggesting that it would make little difference if transcriptome identities were mis-assigned between these cell types. The current understanding of these cells' developmental potential supports the notion that they should be difficult to distinguish from each other. For example, the sister cells ABa and ABp of the 4-cell stage are initially developmentally equivalent (Priess and Thomson, 1987). In the future, if features are identified that more clearly differentiate these cell types, our existing single-cell transcriptomes can be revisited with those features in mind.

5.6 A Stark Contrast in mRNA Composition between Germ Cell Precursors and Somatic Cells

One pattern that is apparent when comparing gene expression across all cell types (Figure 2.4B) is that there is a prevalent distinction between the mRNA composition of the somatic cells and the germ cells (including the somatic sister of each germ cell precursor). Previous studies, such as Seydoux and Fire (1994), have observed this contrast in transcript composition between the germ and soma. Their reliance on *in situ* hybridization necessarily restricted the number of

such genes they were able to study (ten genes), whereas Tintori et al. (2016) expanded their findings to thousands of genes. Differences between ‘‘immortal’’ germ cells and ‘‘mortal’’ somatic cells have fascinated researchers for over a century (Weismann, 1893; Boveri, 1910; Schierenberg and Strome, 1992; Lai and King, 2013; Lehmann and Ephrussi, 2007; Yamanaka, 2007). The datasets described in this document quantitatively identify thousands of genes with differential transcript abundances between the germ and soma. Furthermore, the datasets includes ‘‘before, during, and after’’ snapshots of somatic descendants of germ cell precursors, in their transition from the germ-like profiles of their parent cell to the somatic profiles of their descendants. This provides a rich view of how a cell's transcriptome changes as it transitions from a germ state to a somatic state over time.

5.7 Cross-Lineage Expression Patterns Highlight Genes that May Share Mechanisms of Gene Regulation

tbx-32 and the five other genes with similar expression patterns are examples of genes whose expression is not continuous from parent to daughter cell, but rather appears in one cell type (EMS) at one stage, then in a different lineage of cells (ABxx) at the next stage. The EMS cell at the 4-cell stage and one of these ABxx cells (ABar) at the 8-cell stage have another feature in common, which is that both orient their mitotic spindles in response to Wnt signaling (Walston et al., 2004). The fact that this specific expression pattern is shared by several genes suggests that a common mechanism may be regulating all of these genes, possibly the previously characterized Wnt signaling. Alternatively, these six genes may play a role in establishing which cells are capable of responding to Wnt signaling.

5.8 Identifying Critical Regulators of Development by Querying Groups of Paralogous Genes

In the work described in Chapter 2, my colleagues and I tested a small subset of genes, and identified two that are critical for embryonic development (Figure 2.6D). These two genes are similar in sequence, and have similar but slightly staggered transcript enrichment patterns (Figure 2.6E). The staggering of these two patterns may represent subfunctionalization after a gene-duplication event. This observation suggests that by considering both paralogy and spatiotemporal transcript abundance, our dataset may reveal patterns about divergence in sequence and function after a gene-duplication event.

In the work described in Chapter 3, again we included paralogy groups alongside single genes in our analyses. We found a group of genes (*zyx-1*, *lim-9* and *plx-1*) encoding proteins with LIM domains, whose collective transcript abundance pattern is correlated with gastrulation. This group of genes may represent gene duplication events of a single gene followed by a specialization of each gene in separate cell types.

5.9 RNAi Targeting *zyx-1* Compromises Gastrulation in the E lineage

The three LIM domain containing proteins described above are broadly conserved, and are best known for actin and myosin organization in thick filament bundling (Hoffman et al. 2006, Kadrmas and Beckerle, 2004). Their proteins have been identified in M-lines, dense bodies and Z-discs in muscles, and sites of actin attachment in focal adhesions (Lecroisey et al., 2013, Qadota et al., 2007, Warner et al., 2011). They have been implicated in focal adhesions, mechanotransduction (Cattaruzza et al., 2004, Hoffman et al., 2012), stretch-induced gene expression (Nix and Beckerle, 1997), planar cell polarity and asymmetric cell division (Wu and Herman, 2006), and have been shown to interact physically with actin cytoskeletal components

such as vinculin and alpha-actinin (Lecroisey et al., 2013). We had considered that these transcripts might be muscle-specific, but two observations lead us to believe that they may be serving morphogenesis-specific functions at this point in embryogenesis. First, *zyx-1* transcripts are enriched in the E cell, whose descendants will only become intestine (Tintori et al., 2016, tintori.bio.unc.edu). Second, these data represent a very early stage of embryogenesis, with transcripts of these three genes detected between 90 and 210 minutes after the first cell division. It is unlikely that muscles are beginning to function this early in embryogenesis. For comparison, *myo-3* and *unc-54*, body wall muscle myosins A and B, are first detected as transcripts at very low levels at 240 minutes, gradually increasing until reaching their peak transcript abundance at 690 minutes (Celniker et al., 2009).

Preliminary studies of *zyx-1*, performed by Goldstein lab masters student Tim Cupp, further support the hypothesis that these LIM containing proteins may be involved in gastrulation. When Tim targeted *zyx-1* by dsRNA injection, embryos showed 17.6%, 28% and 11.5% embryonic lethality across three replicates with sample sizes of 34, 25, and 78 embryos, respectively (Cupp, 2017). By comparison, mothers injected with mock dsRNA showed embryonic lethality of 1.3%, 1.4%, and 0% (n=147,72,102). In typical embryos, E descendants internalize before dividing from two to four cells. When Tim targeted *zyx-1* transcripts by dsRNA, E descendants in 25% of embryos did not internalize before dividing into four cells (n=25).

During apical constriction of the two internalizing E descendants in typical embryos, myosin movement begins before the cell membrane begins to constrict. This creates two distinct phases of apical constriction— before the membrane becomes linked to the contracting myosin, and after (Figure 5.1A). In embryos in which Tim targeted *zyx-1* by dsRNA, the second phase of

apical constriction showed much higher slippage between myosin and membrane than in wild type, suggesting that contracting myosin was not effectively linking to the cell membrane (Figure 5.1B)(Cupp, 2017). Collectively, this data suggests that in E descendants, *zyx-1* may be important for linking contracting myosin to the cell membrane.

5.10 An Interactive Data Visualization Tool To Share Genomic Data With Non-Genomic Scientists

Data sharing is a tenet of the scientific community. I wanted to ensure that the data generated by this research was not only *available* to other researchers, but was also *accessible* to them. In the not-too-distant future, computational skills will be requisite in every researcher's toolkit, but at the moment there are many who study the very cells that we have generated transcriptomes for, but who are not likely to mine our sequencing data themselves. This inaccessibility barrier between computational and non-computational biologists harms both sides— non-computational biologists cannot use the data, and therefore undervalue the work done by computational biologists. The same instincts that drive me to bridge communication gaps between scientists and others in the future (as described in Chapter 4), drove me to attempt to bridge this gap between computational and non-computational biologists with my own dataset. My colleagues and I designed an online interactive data visualization tool of all the transcriptome datasets from the 1- to 16-cell stage. This tool allows users to import lists of genes and browse their transcript abundance patterns, or to define abundance patterns of interest and discover new genes that match the pattern.

FIGURES

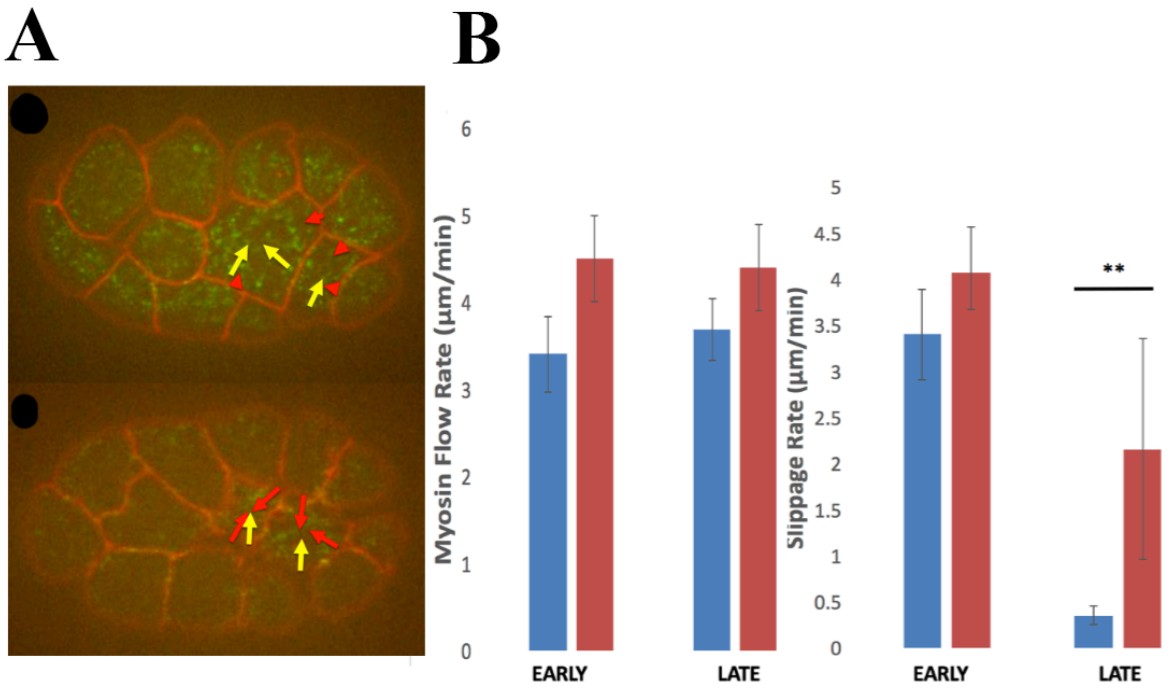

Figure 1 - Embryos treated with *zyx-1* dsRNA show decreased myosin-membrane linkage. *Figure from Tim Cupp's master's thesis*

a. Top: At the onset of apical constriction on the E cells, myosin (green) flows from the lateral edges of the cell surface inward, while cell membrane position (red) remains unchanged. Bottom: During the second phase of apical constriction, lateral cell membrane moves towards the center of the apical surface in concert with myosin.

b. In embryos injected with *zyx-1* dsRNA (red bars), myosin flow and membrane slippage are similar to that of mock-injected embryos (blue bars) during the early phase, but slippage is significantly higher during the later phase.

REFERENCES

Achim, K., Pettit, J.-B., Saraiva, L. R., Gavriouchkina, D., Larsson, T., Arendt, D., & Marioni, J. C. (2015). High-throughput spatial mapping of single-cell RNA-seq data to tissue of origin. Nature Biotechnology, 33(5), 503–509. http://doi.org/10.1038/nbt.3209

Baugh, L. R., Hill, A. A., Claggett, J. M., Hill-Harfe, K., Wen, J. C., Slonim, D. K., et al. (2005). The homeodomain protein PAL-1 specifies a lineage-specific regulatory network in the C. elegans embryo. Development (Cambridge, England), 132(8), 1843–1854. http://doi.org/10.1242/dev.01782

Baugh, L. R., Hill, A. A., Slonim, D. K., Brown, E. L., & Hunter, C. P. (2003). Composition and dynamics of the Caenorhabditis elegans early embryonic transcriptome. Development (Cambridge, England), 130(5), 889–900.

Biase, F. H., Cao, X., & Zhong, S. (2014). Cell fate inclination within 2-cell and 4-cell mouse embryos revealed by single-cell RNA sequencing. Genome Research, 24(11), 1787–1796. http://doi.org/10.1101/gr.177725.114

Boveri, T. (1910). Über die Teilung centrifugierter Eier von Ascaris megalocephala. Archiv Für Entwicklungsmechanik Der Organismen, 30(2), 101–125. http://doi.org/10.1007/BF02263806

Cao, J., Packer, J.S., Ramani, V., Cusanovich, D.A., Huynh, C. and Daza, R. (2017). Comprehensive single-cell transcriptional profiling of a multicellular organism. Science. 357, Issue 6352, pp. 661-667. http://doi.org/10.1126/science.aam8940

Cattaruzza M, Lattrich C, Hecker M (2004). Focal adhesion protein zyxin is a mechanosensitive modulator of gene expression in vascular smooth muscle cells. Hypertension 43, 726–730.

Celniker, S.E., Dillon, L.A., Gerstein, M.B., Gunsalus, K.C., Henikoff, S., Karpen, G.H., Kellis, M., Lai, E.C., Lieb, J.D., MacAlpine, D.M., Micklem, G., Piano, F., Snyder, M., Stein, L., White, K.P. and Waterston, R.H. (2009). modENCODE Consortium. Unlocking the secrets of the genome. Nature. 459(7249):927-30. http://doi.org/10.1038/459927a.

Gilmour, D., Rembold, M. & Leptin, M. (2017). From morphogen to morphogenesis and back. Nature. 541(7637):311-320. http://doi.org/10.1038/nature21348

Grün, D., Lyubimova, A., Kester, L., Wiebrands, K., Basak, O., Sasaki, N., et al. (2015). Single-cell messenger RNA sequencing reveals rare intestinal cell types. Nature, 525(7568), 251–255. http://doi.org/10.1038/nature14966

Hashimshony, T., Feder, M., Levin, M., Hall, B. K., & Yanai, I. (2015). Spatiotemporal transcriptomics reveals the evolutionary history of the endoderm germ layer. Nature, 519(7542), 219–222. http://doi.org/10.1038/nature13996

Hashimshony, T., Wagner, F., Sher, N., & Yanai, I. (2012). CEL-Seq: single-cell RNA-Seq by multiplexed linear amplification. Cell Reports, 2(3), 666–673. http://doi.org/10.1016/j.celrep.2012.08.003

Hoffman, L.M., Jensen, C.C., Chaturvedi, A., Yoshigi, M. and Beckerle, M.C. (2012). Stretch-induced actin remodeling requires targeting of zyxin to stress fibers and recruitment of actin regulators. Mol Biol Cell 23, 1846–1859.

Jaitin, D. A., Kenigsberg, E., Keren-Shaul, H., Elefant, N., Paul, F., Zaretsky, I., et al. (2014). Massively Parallel Single-Cell RNA-Seq for Marker-Free Decomposition of Tissues into Cell Types. Science (New York, N.Y.), 343(6172), 776–779. http://doi.org/10.1126/science.1247651

Kadrmas, J.L. and Beckerle, M.C. (2004). The LIM domain: from the cytoskeleton to the nucleus. Nat Rev Mol Cell Biol 5, 920–931.

Lai, F., & King, M. L. (2013). Repressive translational control in germ cells. Molecular Reproduction and Development, 80(8), 665–676. http://doi.org/10.1002/mrd.22161

Lecroisey, C., Brouilly, N., Qadota, H., Mariol, M., Rochette, N. C., Martin, E., Benian, G. M., Segalat, L. S., Mounier, N. & Gieseler, K. (2013). ZYX-1, the unique zyxin protein of Caenorhabditis elegans, is involved in dystrophin-dependent muscle degeneration.. Mol Biol Cell, 24, 1232-49. http://doi.org/10.1091/mbc.E12-09-0679

Lehmann, R., & Ephrussi, A. (2007). Germ Plasm Formation and Germ Cell Determination in Drosophila. In Ciba Foundation Symposium 182 - Germline Development (pp. 282–304). Chichester, UK: John Wiley & Sons, Ltd. http://doi.org/10.1002/9780470514573.ch16

Levin, M., Hashimshony, T., Wagner, F., & Yanai, I. (2012). Developmental Milestones Punctuate Gene Expression in the Caenorhabditis Embryo. Developmental Cell, 22(5), 1101–1108. http://doi.org/10.1016/j.devcel.2012.04.004

Lynch, M. and Conery, J.S. (2000). The evolutionary fate and consequences of duplicate genes. Science. 290:1151–1155

Nix, D.A., Beckerle, M.C. (1997). Nuclear–cytoplasmic shuttling of the focal con- tact protein, zyxin: a potential mechanism for communication between sites of cell adhesion and the nucleus. J Cell Biol 138, 1139–1147.

Pollen, A. A., Nowakowski, T. J., Shuga, J., Wang, X., Leyrat, A. A., Lui, J. H., et al. (2014). Low-coverage single-cell mRNA sequencing reveals cellular heterogeneity and activated signaling pathways in developing cerebral cortex. Nature Biotechnology, 32(10), 1053–1058. http://doi.org/10.1038/nbt.2967

Priess, J. R., & Thomson, J. N. (1987). Cellular interactions in early C. elegans embryos. Cell, 48(2), 241–250.

Qadota, H., Mercer, K. B., Miller, R. K., Kaibuchi, K. & Benian, G. M. (2007). Two LIM domain proteins and UNC-96 link UNC-97/pinch to myosin thick filaments in Caenorhabditis elegans muscle.. Mol Biol Cell, 18, 4317-26. http://doi.org/10.1091/mbc.E07-03-0278.

Satija, R., Farrell, J. A., Gennert, D., Schier, A. F., & Regev, A. (2015). Spatial reconstruction of single-cell gene expression data. Nature Biotechnology, 33(5), 495–502. http://doi.org/10.1038/nbt.3192

Sawyer, J. M., Glass, S., Li, T., Shemer, G., White, N. D., Starostina, N. G., Kipreos, E.T., Jones, C.D. & Goldstein, B.(2011). Overcoming redundancy: an RNAi enhancer screen for morphogenesis genes in Caenorhabditis elegans. Genetics, 188(3), 549–564. http://doi.org/10.1534/genetics.111.129486

Schierenberg, E., & Strome, S. (1992). The establishment of embryonic axes and determination of cell fates in embryos of the nematode Caenorhabditis elegans. Seminars in Developmental Biology, 3, 25–33. http://doi.org/10.1234/12345678

Sulston, J. E., Schierenberg, E., White, J. G., & Thomson, J. N. (1983). The embryonic cell lineage of the nematode Caenorhabditis elegans. Developmental Biology, 100(1), 64–119.

Tintori SC, Osborne Nishimura E, Golden P, Lieb JD, Goldstein B. (2016). A Transcriptional Lineage of the Early C. elegans Embryo. Dev Cell. 38(4):430-44. http://doi.org/10.1016/j.devcel.2016.07.025.

Trapnell, C., Cacchiarelli, D., Grimsby, J., Pokharel, P., Li, S., Morse, M., et al. (2014). The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. Nature Biotechnology, 32(4), 381–386. http://doi.org/10.1038/nbt.2859

Walston, T., Tuskey, C., Edgar, L., Hawkins, N., Ellis, G., Bowerman, B., et al. (2004). Multiple Wnt signaling pathways converge to orient the mitotic spindle in early C. elegans embryos. Developmental Cell, 7(6), 831–841. http://doi.org/10.1016/j.devcel.2004.10.008

Warner, A.D., Qadota, H., Benian, G.M., Vogl, A. & Moerman, D. G. (2011). The Caenorhabditis elegans paxillin orthologue, PXL-1, is required for pharyngeal muscle contraction and for viability. Mol Biol Cell, 22, 2551-63. http://doi.org/10.1091/mbc.E10-12-0941

Weismann, A. (1893). The germ-plasm. London: W. Scott.

Wieschaus, E. (1997). Nobel Lecture. In N. Ringertz (Ed.), Nobel Lectures. Singapore.

Woollard, A. (2005). Gene duplications and genetic redundancy in C. elegans. WormBook, ed. The C. elegans Research Community, WormBook. http://doi.org/10.1895/wormbook.1.2.1.

WormBase web site, http://www.wormbase.org, release WS260, August 2017.

Wu, M., and Herman, M.A. (2006). A novel noncanonical Wnt pathway is involved in the regulation of the asymmetric B cell division in C. elegans.. Dev Biol, 293, 316-29. http://doi.org/10.1016/j.ydbio.2005.12.024.

Xue, Z., Huang, K., Cai, C., Cai, L., Jiang, C.-Y., Feng, Y., et al. (2013). Genetic programs in human and mouse early embryos revealed by single-cell RNA sequencing. Nature, 500(7464), 593–597. http://doi.org/10.1038/nature12364

Yamanaka, A., Yada, M., Imaki, H., Koga, M., Ohshima, Y., & Nakayama, K.-I. (2002). Multiple Skp1-related proteins in Caenorhabditis elegans: diverse patterns of interaction with Cullins and F-box proteins. Current Biology : CB, 12(4), 267–275.

Yan, L., Yang, M., Guo, H., Yang, L., Wu, J., Li, R., et al. (2013). Single-cell RNA-Seq profiling of human preimplantation embryos and embryonic stem cells. Nature Structural & Molecular Biology, 20(9), 1131–1139. http://doi.org/10.1038/nsmb.2660

Zeisel, A., Muñoz-Manchado, A. B., Codeluppi, S., Lönnerberg, P., La Manno, G., Juréus, A., et al. (2015). Brain structure. Cell types in the mouse cortex and hippocampus revealed by single-cell RNA-seq. Science (New York, N.Y.), 347(6226), 1138–1142. http://doi.org/10.1126/science.aaa1934