Title

Exploring the regulation of gene expression in the early Drosophila embryo using genetics and single-nucleus RNA-sequencing

Author

Albright, Ashley

Publication Date

2021

Peer reviewed|Thesis/dissertation

Exploring the regulation of gene expression in the early *Drosophila* embryo using genetics and single-nucleus RNA-sequencing

By

Ashley Renee Albright

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Molecular and Cell Biology

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Michael B. Eisen, Chair
Professor David Bilder
Professor Xavier Darzacq
Professor Louise Glass

Summer 2021

Exploring the regulation of gene expression in the early *Drosophila* embryo using genetics and single-nucleus RNA-sequencing

Ashley Renee Albright

Abstract

Exploring the regulation of gene expression in the early *Drosophila* embryo using genetics and single-nucleus RNA-sequencing

By

Ashley Renee Albright

Doctor of Philosophy in Molecular and Cell Biology

University of California, Berkeley

Professor Michael Eisen, Chair

Gene expression is primarily regulated by various genomic elements, namely enhancers. Enhancers are stretches of non-coding DNA that influence gene expression across space and time. Much of what we understand about enhancers comes from the study of spatially patterned gene expression and transcription factor localization. We know that enhancers contain clusters of transcription factor binding sites, but these sites alone do not define the identity of an enhancer and the question of how enhancer identity is specified remains unanswered. The work that I describe in this dissertation aims to explore other factors associated with enhancer activity in the early *Drosophila melanogaster* embryo and to improve existing technologies for the study of gene expression. In Chapter 1, I describe the many factors associated with enhancer activity as well as a new technology that may allow us to finally answer questions surrounding enhancer identity.

Transcriptional co-factors, such as histone modifiers or insulator proteins, either directly or indirectly influence transcription factor binding and enhancer activity, which can drastically affect patterned gene expression. In Chapter 2, I screened for transcriptional co-factors with enhancer-specific defects by characterizing the effects of reduced transcriptional co-factor expression on enhancer activity. Given the importance of enhancer activity in regulating patterned gene expression, I assayed the expression of a classic patterned gene as a proxy for enhancer activity and the expression of a non-patterned gene as a transcriptional control. I found several candidates, all histone modifiers, that specifically affect expression of a patterned gene while the non-patterned gene expression remained unchanged. Ultimately, I found a wide and overlapping variability in the extent of knockdown of these co-factors and I pursued other methods as a result. Nonetheless these data suggest that a system outside of the canonical view of transcription factor binding exists to define enhancer activity and identity.

In early development, enhancer activity is primarily driven by maternally-deposited RNAs and proteins prior to zygotic genome activation. One of the largest hurdles in understanding the regulation of zygotic gene expression at this early point in development lies in our ability to distinguish between maternal and zygotic RNAs. Throughout the course of my PhD, single-cell RNA-sequencing became an increasingly popular way to study gene expression on a smaller scale than ever before. With that being said, zygotic genome activation occurs

concurrently with cellularization in the early *Drosophila melanogaster* embryo, thus isolation of cells is not possible. In Chapter 3, I describe a set of experiments and analyses demonstrating the use of single-nucleus RNA-sequencing in the early *Drosophila melanogaster* embryo using an insulator protein as a case study. Under the assumption that the vast majority of RNAs present in the nucleus are zygotic transcripts, this allows us to assay zygotic gene expression outside of the context of cytoplasmic and maternal RNAs to really examine defects in transcription. I found that the nuclei retain spatial information, or where the nuclei were located in the embryo prior to dissociation. I also found examples of patterned genes that are differentially expressed upon the loss of the insulator protein in individual clusters, but not in bulk. These results highlight the importance of establishing the use of single-nucleus RNA-sequencing. From this, we now have the capacity to understand global changes in spatial gene expression across the embryo, giving us the ability to answer questions regarding the regulation of patterned gene expression.

As discussed above, single-nucleus RNA-sequencing is a powerful tool; however, current technologies are limited either by the number of nuclei captured or the accessibility of the technique in the first place. In an effort to improve barcoding of individual cells or nuclei, I began a collaboration with a group of researchers. In Chapter 4, I discuss testing a set of catalytic DNA hairpin oligos as a means of barcoding cDNA in individual cells or nuclei. I demonstrated that these hairpins are highly efficient and specific in barcoding single-stranded DNA molecules *in vitro*. I show additional work towards optimizing this reaction following reverse transcription; however, a completed method is outside of the scope of this dissertation and this work is ongoing.

Altogether, the work that I have completed throughout this dissertation has progressed our understanding of the regulation of gene expression by working outside of the canonical view of transcription, bringing new technologies to the table, and potentially drive the field forward with the development of new methods to improve existing technologies.

Dedication

To my mom,
for encouraging me to pursue my dreams,
inspiring me to become by best self,
and without whom I would never have gotten this far.

Acknowledgements

I never thought coming from a small town that I would move across the country to pursue my dreams and obtain a PhD. I would not be where I am today if it weren't for the many people that helped me along the way. I would like to thank my friends, family, mentors, and dogs for getting me this far.

First, I would like to thank my advisor, Mike Eisen, for supporting me for the last six years. I always joke that I am turning into you, given the amount of profanity that I use, the sports that I yell about, and the disdain I have for academic bureaucracy. In reality however, you have given me the space to become myself. You believed in me when I didn't, gave me the courage to stand up for myself, and made me realize the power that I had to make academia a better place. I'll never forget that, even though your sports teams are garbage and you have difficulty in managing your calendar at times (but you know I appreciate the work that you do for science).

I would also like to thank my thesis committee, David Bilder, Xavier Darzacq, and Louise Glass. David – thank you for your support over the years in both science and personal matters, you were an excellent Head Graduate Advisor and I am forever grateful for your guidance. Xavier – thank you for the excellent feedback that you have provided over the years as well as the experience I had teaching alongside you, I truly appreciate all of the advice that you have given me. Louise – thank you for providing a different perspective on my work and the helpful discussions, your input was always immensely helpful.

I would also like to thank Elçin Ünal, Diana Bautista, and Don Rio for your encouragement and support. You all have gone above and beyond to help me throughout this process.

A huge thank you to all of my labmates, past and present. Jenna – I also would not have gotten through grad school without you and I am so glad that we experienced everything together. You were there for me in the brightest and darkest times and I will be forever grateful for everything that you have done for me. Also, I still love you even though you spelled my name wrong in your dissertation acknowledgements section. Stadler – I can't believe our families are from the same small town and it was really nice having someone that understood where I came from literally right next to me every day. You have also done so much to support my work and I am so grateful. Colleen – you have been such a wonderful friend and labmate. Thank you for all of the music trivia outings, for answering all of the fly and science questions that I asked on a regular basis without judgment, as well as the unwavering support towards the end of my PhD. Marc – thank you for answering all of my computational and statistics questions, and for your entertaining snarky banter. I feel like I laughed at least once a day because of you, bringing me much joy. Ciera – you are undoubtedly the coolest person that I have ever met, I like that we bonded over stationery, and you have inspired me to constantly strive to make my data reproducible and usable. Without you, I also never would have found about the opportunity to go to Tokyo for a few weeks which was an amazing experience. Victoria and Laura – I feel like we didn't to spend enough time together in person but in the time that we were together I always felt supported, we had some good laughs, and I am so glad to know the both of you. Holli – thank

you so much for keeping the lab organized (despite some of our inability to do so) and for getting things done. Augusto and Xiao-yong – thank you for all of your help and feedback on my work.

Alli, Carolyn, and Elizabeth – in addition to Jenna, you all were my original support system when I joined the lab. You all helped me get my PhD started, and I couldn't have done it without you. You all taught me so much about flies, molecular biology, and most importantly, resilience.

Thank you to my mentee, Ruchika Singla, for helping me with experiments and providing me with invaluable experience on my journey of becoming a better mentor and leader.

Thank you to my cohort – especially Helen, Grant, and Jenna. I cherish my best and my worst days, because you all (and boba or food) were always there for me. You all are my best friends for life and made me feel welcome and feel like I belonged in graduate school. I am here for anything that any of you ever need.

Thank you to my research advisor at UNC – Greg, I started working in your lab almost 10 years prior to my writing this dissertation and I am forever grateful for your continuing support. When I joined your lab, I didn't really know what working in a lab was like and didn't know that a PhD was even in the cards. I am so glad that I half-jokingly asked you to fly to UC Davis. As we drove by the exit to UC Berkeley on our way, I couldn't believe that you suggested that I apply to Berkeley for graduate school. I never thought I'd make it, but you did and still do. I can't tell you how much I appreciate that.

Thank you to my best friends from UNC – Raleigh, Becca, Peter, and Rodrigo. Raleigh and Rodrigo, I am so fortunate that you both lived so close to me during graduate school. I am so grateful to have such wonderful friends to enjoy life with, as well as escape the stresses of graduate school. Becca and Peter, I am so grateful for your continued support and I can't wait for us all to be reunited together!

To my family – Mom, Dad, Grandmama, Brooke, Kevin, Zach, Mammy, thank you so much for getting me where I am today. I'm sorry I moved so far away, but each time we saw each other or talked, I was reminded of the important things in life and suddenly all of the stress that accompanied graduate school disappeared.

Finally, I would like to thank my life partner, Cole. You love me when I hate myself, you are there when I cry, you make me laugh even when I don't want to, you are the reason that I have reached the finish line. I am so glad that you took a chance when you moved across the country to be with me on this journey and I cannot wait to spend the rest of my life with you. To Cole's family – I am so grateful for you all welcoming me into your family with open arms, and for your continued support.

Super finally, thank you so much to the best dog on the entire planet, Barney for his unconditional love. Nothing will ever make me happier than a happy dog waiting at home after a long day in lab.

Table of Contents

Chapter 1: Introduction

Exactly 100 years before the start of my PhD, Thomas Hunt Morgan and his colleagues showed that chromosomes contain hereditary information passed from generation to generation by studying various traits in thousands of *Drosophila melanogaster,* colloquially known as fruit flies[1]. Years later, Rosalind Franklin's work allowed others to determine the structure of DNA[2], the material that encodes information contained in chromosomes. Since then we have learned that our genome contains genes that are transcribed and translated into proteins. Proteins make up many different structures in many different cell types across all walks of life. Gene expression and its regulation are essential to understand why we, and every other organism in the universe, are the way we are, but also why things can go wrong. As an example, when a gene normally expressed in rapidly-growing cells, like skin, is wrongly turned on in slow-growing tissues like the brain, this can make too much of a protein that increases chances of cancer. Even though over 100 years have passed since the discovery of genetic material in fruit flies, many questions surrounding the regulation of gene expression remain unanswered.

Enhancers regulate gene expression level and patterns across time

The regulation of gene expression is essential to the survival of every living organism. Enhancers, or stretches of non-coding DNA regulating the transcription of nearby genes, in particular are largely responsible for controlling precise spatiotemporal gene expression over the course of development. Despite being first described more than 40 years ago[3], we do not fully understand enhancer function. Many early contributions towards our understanding of patterned gene expression, along with the discovery of genetic material in the first place, were carried out in early embryonic development of fruit flies, specifically *Drosophila melanogaster*. The most notable study of patterned gene expression led to the Nobel Prize in Physiology or Medicine in 1995. Together, Christiane Nüsslein-Volhard and Eric Wieschaus conducted a genetic screen and found 15 loci that affected larval segmentation pattern[4]. The other 1/3rd of the prize awarded that year was given to Edward Lewis for his discovery of eight genes that also affect segmentation[5].

While they did not necessarily make the connection to enhancers straight away, they found many genes that significantly affect patterning and segmentation in early development. The expression pattern of those genes, we now know, is primarily governed by enhancer activity[6,7].

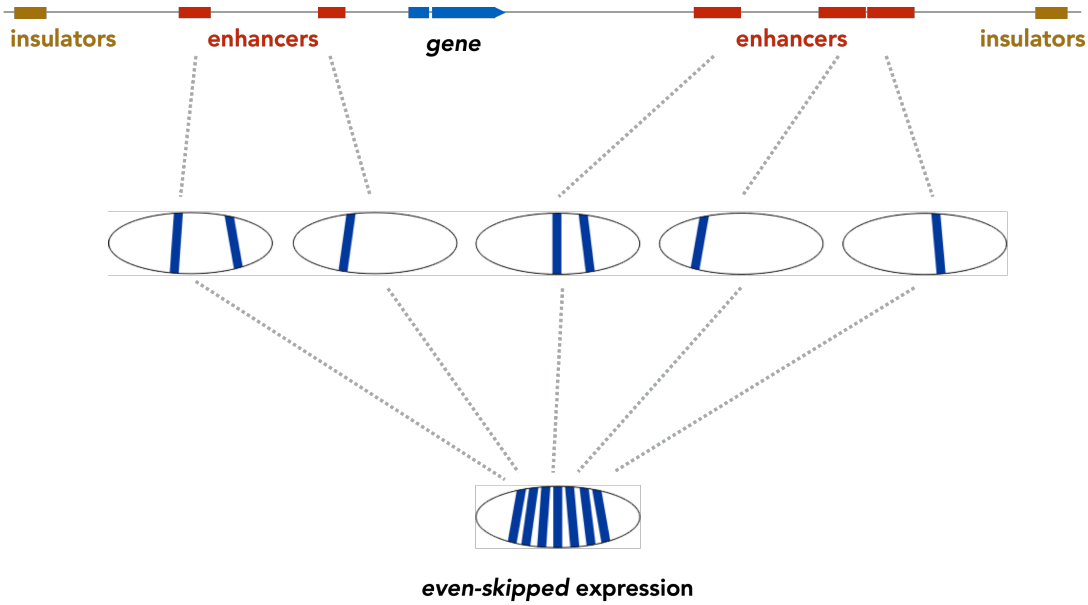

Figure 1.1 Diagram of regulatory elements that lead to *even-skipped* expression Figure depicts a simplified overview of the regulatory elements surrounding *even-skipped*. Insulators (yellow) and enhancers (red) somehow work together in order to coordinate gene expression (blue) in the right place and the right time. Image courtesy of Mike Eisen.

Building on earlier work on fruit fly development, much of our understanding the regulation of gene expression by enhancers was also conducted in the early *Drosophila melanogaster* embryo, focusing on *even-skipped* enhancers. *Even-skipped* was actually one of the genes found to affect patterning in Christiane Nusslein-Volhard and Eric Wieschaus's famous genetic screen. Eventually, we came to know that five enhancers placed in different locations surrounding the *even-skipped* gene somehow coordinate with one another and the *even-skipped* promoter to consistently give rise to seven distinct stripes at a particular time of development in every single developing embryo (see Figure 1.1).

Further work showed that enhancer activity occurs via the coordinated binding of multiple transcription factors. Some transcription factors act as activators to promote enhancer activity, some as repressors to lessen enhancer activity, and some can act as both depending on the context. In the end, many different factors have to work together to give rise to the correct pattern[8,9]. Although the early genetic screens for patterning defects were not conducted in the context of regulatory regions, like enhancers, nonetheless those early screens defined the importance of patterned gene expression in early development.

Enhancer mutations drive evolution and disease

In addition to coordinating development, alterations of enhancer function as a result of sequence divergence have long been purported as major drivers of evolution as well[10,11], which has been substantiated by many recent studies. As an example, pelvic reduction in threespine sticklebacks has been linked to differential expression of *Pitx1,* in two isolated populations studied. This suggests that changes in cis-regulatory elements may have occurred[12], and mutations in a *Pitx1* enhancer were later linked to the evolution of pelvic reduction[13]. Pelvic reduction is important to

note, as it has implications in evolutionary adaptations to different environments and this gene functions in human limb development as well. Understanding the molecular mechanisms that specify enhancer activity is critical for furthering our comprehension of the role that enhancers have in evolution.

There are also strong implications for changes in cis-regulatory elements in disease. The vast majority (in some cases over 90%) of hits found in Genome Wide Association Studies (GWAS) of many complex diseases ranging from Alzheimer's to Type II Diabetes and autoimmune diseases are found in non-coding regions, including enhancers. These findings suggest that changes in regulatory regions, including enhancers, might be a primary cause of many diseases[14–16]. In the case of pre-axial polydactyly, which manifests as extra digits on the radial (thumb) side of the hand, a single base pair change in a Sonic hedgehog limb enhancer has been shown to be responsible for the disease phenotype[17]. As we continue to uncover additional associations between enhancers and disease and show that some of these associations are causal, it is essential that we understand how enhancer identity is defined in the first place.

Transcription factor binding sites do not define an enhancer

In the late 1990s and early 2000s, the genome sequencing craze was just getting started. People thought that we could solve everything about gene regulation from genomes alone. The first sequenced genome was of a bacteria, *Haemophilus influenzae,* with a size of 1.8 million base pairs[18]. In contrast, the *Drosophila melanogaster* genome contains roughly 137 million base pairs, and the human genome contains roughly 2.8 billion base pairs[19]. As a model system, *Drosophila melanogaster* benefits from a long-standing body of work more than 100 years long, and benefits from having a much smaller genome to study. The exact composition of this genome relative to other eukaryotic genomes, such as humans, may differ; however, many principles of genome structure and the central dogma of biology remain the same. DNA is transcribed into RNA, and RNA is translated into proteins across all walks of life. We can learn much about the regulation of gene expression from the *Drosophila melanogaster* genome, which benefits from its much smaller size, and apply this knowledge to human gene regulation as well.

At the time the first sequenced *Drosophila melanogaster* genome was released[20], people began to question whether we could use knowledge of transcription factor binding motifs to predict the presence of a regulatory element, such as an enhancer[21–23]. While this approach did generate novel enhancer candidates, this still does not explain how a particular cluster of transcription factor binding sites becomes an enhancer. Bicoid, Caudal, Knirps, Krüppel, and Hunchback are just a few examples of *Drosophila* transcription factors that regulate transcription and patterning of many early genes, but the binding sites for each of these are scattered all throughout the genome[21].

If we take an imaginary transcription factor with a 6 base long motif and a genome size of 139.5 million bases in *Drosophila*, this motif would be found (1/4)^6 * 139.5 million, or roughly 34,000 times. Bicoid has a primarily 6 base long motif, with so many possible binding sites how does Bicoid or any transcription factor for that matter decide which binding sites to access? Even 20 years after the first release of the *Drosophila* genome, many of these questions remain unanswered.

Remarkably, enhancers retain their original function after sequence divergence, a difference between related DNA sequences, when transcription factor binding sites are significantly altered. Hare et al (2008) showed that the expression patterns of multiple early developmental genes are conserved between *Drosophila melanogaster* and *Themira minor,* a distantly related species of fly. They also showed evidence of minimal sequence conservation between the two in both enhancer and non-enhancer regions surrounding *eve*[24]. Thus, transcription factor binding sites in enhancers can be gained, lost, or altered over evolutionary time, yet the function and the pattern of the enhancer remains the same. Indeed, others have shown that functional binding site turnover, or a case where an added functional binding site allows for loss of another, is widespread in the *Drosophila* lineage[25].

Interestingly, Hare et al (2008) also showed using enhancer-reporter constructs integrated into the *Drosophila melanogaster* genome that enhancers from several distantly related fly species, *Sepsis cynipsea, Themira putris,* and *Themira superba,* replicate nearly exactly the same pattern produced by *Drosophila melanogaster*[24]. Despite extensive binding site turnover, others have also shown that the level of gene expression is also highly conserved[26]. The fact that these distant sequences recapitulate the expression pattern and level of the corresponding gene in *Drosophila melanogaster,* suggests that DNA sequence alone, including transcription factor binding site motifs, cannot define enhancer identity. Intriguingly, intervening sequences between transcription factor binding sites have also been shown to play a role in transcription factor binding and enhancer function[27–29], but the mechanism by which this occurs is also not clear. Transcription factor binding sites and intervening sequences are certainly important in terms of enhancer function; however, the presence of binding sites all over the genome and conservation of patterning in different sequences suggests that something else must define enhancer identity.

Changes in chromatin accessibility are associated with enhancer function

Genomes are not just composed of naked DNA, instead genomes are covered in RNA and proteins that are packaged in a condensed matter, known as chromatin. With recent advances in technology, studies have shown that accessible chromatin is a critical feature of an active enhancer. ATAC-Seq, or Assay for Transposase Accessible Chromatin using sequencing, allows us to rapidly examine chromatin accessibility with minimal sample input[30]. Researchers have shown using ATAC-Seq that formation of regions of accessible chromatin at enhancers throughout the development of the early *Drosophila melanogaster* embryo occurs with precise timing and correlates with the presence of specific transcription factor binding sites[31]. My colleague (and best friend), Jenna Haines, also showed by slicing embryos in half that although early enhancers are generally accessible across the entire embryo, accessibility tends to be higher in the region where the enhancer is active[32]. A similar case may be made for enhancer accessibility in differentiating mouse embryonic stem cells, as intragenic regulatory element accessibility dynamically changed throughout differentiation[33]. With similar changes occurring in other eukaryotes, our knowledge of chromatin accessibility dynamics in the developing *Drosophila melanogaster* embryo allows us to understand the regulation of gene expression in many organisms.

Given that chromatin accessibility at enhancers correlates with their position within the embryo, as well as with specific transcription factor binding sites, establishment of open chromatin likely occurs in part due to transcription factor binding. Another colleague, Colleen Hannon, has shown that an anterior patterned transcription factor, Bicoid, affects chromatin accessibility in regions of high concentration, but not at regions of lower concentration where binding sites are already accessible. Hannon et al (2017) also showed that Bicoid targets in regions of high concentration have a higher nucleosome occupancy, which suggests that Bicoid itself may influence the chromatin environment[34].

Interestingly, enhancer accessibility is also thought to play a role in evolution based on evidence of a new enhancer co-opting an accessible sequence from the ancestral enhancer[35]. Although the connection between chromatin accessibility, enhancer activity, and possibly evolution as well has been established, accessibility still does not define an enhancer as many other regions of the genome require accessibility for function as well.

Chromatin modifications are associated with transcriptional activity

Often, transcription factors interact with non-transcription factor proteins, or co-factors, in order to carry out their respective functions. Interestingly, a study in yeast showed that a transcriptional co-factor does not contain a DNA-binding region itself, yet interacts with a transcription factor and is necessary for this factor to recognize one of its binding site motifs[36]. Many modes of regulating transcription factor binding exist (reviewed in [37]), but the fact that in some cases of transcription factor binding cannot occur without a co-factor has clear implications on enhancer activity and identity. Chromatin modifiers, histone modifiers in particular, comprise a major group of transcriptional co-factors.

In many organisms, enhancers are typically associated with the presence of certain histone marks while promoters and other genomic elements are associated with a different set of marks (reviewed in [38]). The idea that chromatin modifications and their respective modifying proteins could be specific to enhancer activity is a promising avenue to explore with regards to the specification of enhancer activity and identity. In the case presented above, a transcription factor requires a co-factor to function, but in other cases transcription factors recruit the chromatin modifiers in order for the modifiers to function[39,40]. The association between chromatin modifications and accessibility is clear, but whether these modifications are a cause or consequence of activity remains unclear. A genome-wide study of several histone marks in the early *Drosophila* embryo shows that the chromatin state is not fully established in early development until zygotic genome activation. In addition, some histone modifications are largely deposited before others[41]. An answer to the question of whether chromatin marks are a cause or consequence of enhancer activity remains elusive, especially with zygotic genome activation and chromatin state establishment occurring at the same time.

Given the association of chromatin marks and the co-factors that modify them with transcription factor binding and enhancer activity, and the dynamic changes in chromatin state over space and

time[31,32,41], chromatin modifiers likely play a role in the specification of enhancer identity. Intriguingly however, one study of a *Drosophila* chromatin modifier specifically associated with enhancer chromatin found that elimination of its enzymatic activity is not necessary for survival and yields minimal phenotypes[42]. This suggests that chromatin modifiers might affect development and enhancer activity through some unknown mechanism, which is another interesting area of exploration surrounding enhancer identity.

Changes in insulator protein binding are associated with chromatin structure and enhancer function

Insulator elements and proteins were initially described for their enhancer-blocking activity when located between an enhancer and a promoter[43–45]. Mammalian insulators are defined by one protein, CCCTC-binding factor (CTCF), whereas *Drosophila* and other arthropods have evolved several insulator binding proteins that bind in different insulator regions across the genome resulting in multiple classes of insulators[46,47]. Aside from their enhancer-blocking activity, insulators have also been shown to play an important role in the physical structure of the genome and the regulation of gene expression.

Mammalian CTCF has been shown to facilitate chromatin looping, when distant regions of the genome come together, by directly interacting with an essential chromosome segregation protein[48]. Mammalian CTCF also requires RNA interactions to form stable loops[49]. The mechanism by which insulators influence genome topology in *Drosophila* is less clear however. Interestingly, very few examples of chromatin looping exist in the early *Drosophila* embryo[50]. Drosophila insulators also rely on several other insulator proteins in addition to Drosophila CTCF (dCTCF)[46,51–53]. In addition to binding at insulators, another *Drosophila* insulator protein, Cp190, also marks active promoters indicating that insulator proteins in general may serve multiple functions[53]. Understanding multiple functions of insulator proteins would be difficult in mammalian systems, as only one insulator protein exists; however, with several redundant insulator proteins in *Drosophila,* we can isolate various functions of each one without causing complete lethality.

Interestingly, despite its importance as the only insulator protein in mammalian cells, proper *dCTCF* expression is not required for embryo or larvae viability and many insulator elements function independently of dCTCF in *Drosophila*[54,55]. Also, removing individual insulator elements from the genome rather than removing an insulator protein can affect chromosome topology, gene expression, or homeotic phenotypes at particular loci in *Drosophila*[56–58]. However, knockdown of different insulator proteins results in minimal gene expression phenotypes[46,52,59,60] (reviewed in [61]). Many of these experiments were conducted in cultured cells, but given that insulators affect patterned gene expression[54,57], it is possible that loss of insulator proteins yield pattern-specific phenotypes not captured with bulk measurements of gene expression in cell culture.

Maternally-deposited RNAs and proteins drive early enhancer function

Early development, including the many features of the regulation of gene expression as discussed above, is governed by maternally-deposited RNAs and proteins in nearly all animals. In *Drosophila melanogaster,* 14 rapid nuclear cycles occur prior to widespread zygotic genome activation and cellularization. While large-scale genome activation does not occur until the 14th nuclear cycle, 59 zygotically expressed genes are detected as early as the 10th cycle[62]. Activation of the zygotic genome coincides with changes in chromatin modifications[41], chromatin accessibility[31], and early enhancer activation (reviewed in [63]). Zygotic genome activation has to occur at some point in all animals, many parallels can be drawn between what we have learned from studying zygotic genome activation in *Drosophila* and doing the same in other animals (reviewed in vertebrates [64]). Because the changes mentioned above occur prior to large-scale zygotic genome activation, prior to a switch in dependency from maternally-deposited to zygotically-expressed gene products, any factor involved in the specification of enhancer identity must be maternally-deposited. This is likely true in other animals besides *Drosophila* as well.

In the early *Drosophila* embryo, a maternally-deposited transcription factor, Zelda, binds to essentially all known enhancers active during the maternal-to-zygotic transition. This binding occurs prior to the known function of these enhancers[65]. Zelda has also been shown to associate with chromatin accessibility and deposition of chromatin marks correlated with active enhancers, suggesting that Zelda influences the chromatin environment as well[40,41]. The same suggestion can be made for other early transcription factors in *Drosophila* and other animals as well[33,34]. The studies uncovering Zelda's role in enhancer activity represent an important progress in our understanding of enhancer function: how they are defined in the genome sequence, and how transcription factors may gain access to the enhancers. With that being said, Zelda binding does not define enhancer identity as it is also essential for promoter function[65].

Technological advancement pushes the boundary our understanding of gene expression

Previous studies mentioned throughout this chapter either assay gene expression globally, without regards to patterning, or examine patterning in a few genes without regards to global gene expression. Understanding effects of transcription factors, chromatin accessibility and modification, and insulator proteins on gene expression in key developmental genes is important in establishing mechanisms of function. However, perturbation effects on individual genes are not sufficient to describe effects on genes across the entire genome. Throughout my PhD, one technology that may solve this particular problem has exponentially increased the number of both cells and genes that we can study under various perturbations.

To my knowledge, the first record of single-cell RNA-sequencing (scRNA-seq) came from single cells manually isolated from mouse blastomeres in 2009[66]. Over the past 12 years, half of which span the duration of my PhD, the number of cells that are included in such experiments has grown exponentially (reviewed in [67]). Despite increased access over the years, currently available single-cell technologies still may not necessarily allow us to answer questions regarding the regulation of gene expression.

For example, commercially available 10x Genomics kits rely on a proprietary microfluidics and specialized equipment. With a single kit, up to 10,000 cells are partitioned into droplets that contain reagents to barcode each cell individually. Commercial availability has the potential to benefit users with convenience, but this requires the purchase and maintenance of expensive equipment. To put this throughput into perspective, a *Drosophila melanogaster* embryo in the 14th nuclear cycle contains on the order of 5,000 to 6,000 nuclei. In the best case scenario, a single kit may yield up to two embryo's worth of information. This is a start, but certainly does not allow for full capture of biological variability in assaying only two individuals.

Recent single-cell methods developed in academic labs have obtained on the order of tens of thousands to over 100,000 cells using combinatorial indexing or split-pool barcoding approaches[68,69]. Combinatorial indexing, or split-pool barcoding approaches utilize combinations of barcodes across 96-well plates over multiple rounds of barcoding (see Figure 1.2 for split-pool barcoding).

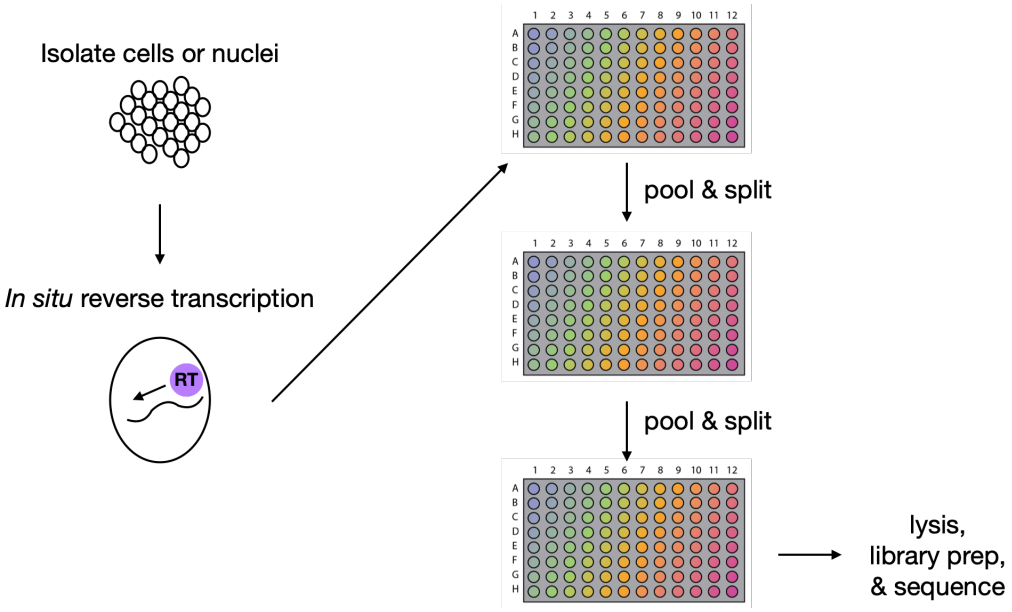

Figure 1.2 Outline of single-cell RNA-sequencing with split-pool barcoding Figure depicts a simplified diagram of single-cell RNA-sequencing by split-pool barcoding. After isolating cells, or nuclei, reverse transcription occurs *in situ* where the cell itself serves a 'container' for the reaction. Unlike microfluidics-based methods, barcoding then occurs in 96 well plates across multiple rounds with each well containing a different barcode, followed by library preparation and sequencing. 96 well plate graphic courtesy of Jenna Haines.

The number of barcoding rounds depends on the number of desired cells sequenced. As an example, three rounds of barcoding across 96 well plates yields 96^3 (884,736) possible combinations of barcodes. Allowing for a 5% clash rate, or the rate at which two cells have the same barcode combination, up to 44,236 cells can be sequenced. Increasing the number of rounds to 4 would yield 96^4 possible barcodes, allowing for up to 4.2 million cells. Barcoding itself can be a cost-prohibitive step in the process, given current split-pool methods are either proprietary or rely on expensive modified oligos that are consumed during barcoding. Developing new barcoding methods that do not rely on proprietary or expensive reagents will

allow us to bring down the cost of single-cell sequencing while increasing the number of genes and cells that we can examine in a single experiment.

If we continue to work towards increasing number of cells (or nuclei) that can be assayed in single-cell RNA-sequencing experiments, we can begin understanding gene expression in many different cell types in several individuals upon some perturbation at the same time. This has been accomplished in the jellyfish, *Clytia hemisphaerica,* by indexing cells from each animal[70]; however, this may be difficult in the early *Drosophila* embryo as no cells exist until after the 14th nuclear cycle. Single-cell RNA-sequencing has been conducted in early *Drosophila melanogaster* embryos after cellularization and zygotic genome activation[71,72]. However to date, single-nucleus RNA-sequencing in pre-cellularization *Drosophila* embryos has not yet been published. The ability to conduct single-nucleus RNA-sequencing in the early *Drosophila* embryo, whether or not we are able to resolve the identity of individual embryos, will allow us to answer questions regarding the regulation of gene expression across different regions of the embryo simultaneously.

Contents of this dissertation

This body of work represents my contributions to our understanding of the regulation of gene expression in the early *Drosophila melanogaster* embryo. First, I describe an RNAi screen for genes that specifically affect enhancer-patterned expression without affecting transcription in general by proxy of a ubiquitously expressed gene (Chapter 2). From this, I found several interesting candidates that alter the expression of a patterned gene, while expression of a ubiquitous gene remains unchanged. However, when quantifying the degree of knockdown in single embryos I revealed an unexpected variability across samples that I published in my first first-author preprint[73].

In order to determine whether or not we can use single-nucleus RNA-sequencing to understand the regulation of gene expression across different regions the embryo, I conducted a series of analyses following single-nucleus RNA-sequencing upon loss of a maternally-deposited insulator protein, dCTCF (Chapter 3). I found after adequate quality control, that nuclei cluster on spatial gene expression in known nucleus types and according to general regions of the embryo. With these data, I demonstrate the possibility that differential expression may occur in specific regions of the embryo that would not be captured by extracting RNA in bulk and sequencing.

Finally, I describe a series of experiments towards the development of a new single-cell RNA-sequencing barcoding strategy. I demonstrate that a catalytic hairpin containing barcodes designed for split-pool barcoding efficiently extends single-stranded DNA *in vitro* (Chapter 4). Although I made significant progress in optimizing this method following reverse transcription, this is an ongoing collaboration that will extend beyond the scope of this dissertation.

I will summarize what I have learned throughout this process and provide my perspective on the future of the regulation of gene expression, among other thoughts in the final chapter of this dissertation (Chapter 5).

Chapter 2: RNAi screen for genes that affect enhancer activity

Abstract

Enhancer activity is essential in establishing early patterns of gene expression in the early *Drosophila melanogaster* embryo. With that being said, the question of how enhancer identity is established remains unanswered. Many factors that influence enhancer activity, other than the presence of transcription factor binding sites, such as transcriptional co-factors, either directly associate with or are at least correlated with the presence of an enhancer in the genome. Even so, the relationship between many co-factors and their resulting direct or indirect effects on enhancer-patterned gene expression is unclear. In this chapter I describe an RNAi screen for genes, primarily co-factors, that affect enhancer activity by conducting a viability assay and dual-color *in situ* hybridizations. Patterning of the *even-skipped* (*eve*) gene serves as a proxy for enhancer-patterned expression, while ubiquitously expressed *sryα* serves as a proxy for non-patterned transcription, or transcription in general. From this screen, I found three candidates (*lid, Gug, nej*), all histone modifiers, that resulted in reduced viability upon knockdown as well as altered or lost *eve* expression with unaffected *sryα* expression. These results suggest that these candidates, either directly or indirectly, could potentially drive enhancer-driven gene expression, but not transcription in general. Determining the extent of knockdown by RNAi in single embryos from one candidate, *lid*, also revealed a wide and overlapping variability of gene expression between control and knockdown embryos. While the exact mechanisms by which these factors impact enhancer identity and activity remain unclear, this work lays a framework in which we think about enhancer identity beyond the canonical view of transcription simply being governed by transcription factors.

Introduction

Transcription factors, transcription factor binding sites, chromatin accessibility, and chromatin modifying proteins are all associated with enhancer activity, but less is known about how each may affect the actual establishment of an enhancer within the genome. Each of these are also important for transcription in general and are not enhancer-specific in any way. Because the zygotic genome is not largely active until the 14th nuclear cycle in the developing *Drosophila melanogaster* embryo anything, including the aforementioned characteristics of enhancer activity, involved in establishing enhancer identity must be maternally-deposited.

In 2004, a group of researchers alongside Christiane Nusslein-Volhard, who conducted the classic 1980 screen for patterning mutants[4], conducted another massive screen for maternal-effect mutations that affecting patterning[74]. Maternal-effect mutations arise in the maternal germline and give rise to an embryonic mutant phenotype irrespective of the zygotic genotype. This screen differs from the 1980 screen in that the method they chose bypassed any potential lethality either in the mother's ovaries or in the zygote itself. From a total of nearly 60,000 crosses they found 45 known and 41 previously undescribed loci. They went on to map many of the new mutants to the genome, but unfortunately over time the undescribed mutants were discarded (personal communication, Stefan Luschnig).

Given the importance of establishing enhancer identity and activity in general, I hypothesize that involved candidates are lethal to either the mother and/or embryo, or specifically lethal in the mother's ovaries. To bypass potential maternal lethality of candidate loss, I decided to screen potential candidates using the maternal-Gal4 shRNA system[75]. With this system, RNAi knockdown occurs in late oogenesis results in a lack of candidate RNAs being passed from mother to embryo. Another technique for bypassing lethality, germline clones, has the benefit of completely eliminating the gene product while allowing the mother to develop normally by relying on FLP-FRT recombinase to generate homozygous mutants during heat shock[76,77]. This approach has been taken before in large-scale genetic screens, such as the maternal-effect screen mentioned above.

In the aforementioned screen, the researchers conducted mutagenesis in an unbiased manner and conducted nearly 60,000 crosses. Germline clones remain a possibility for use in follow-up experiments, but I chose the maternal-Gal4 shRNA system for screening purposes as it is the more directed and faster approach. In designing my own screen for genes that affect specifically affect enhancer activity, it made sense to limit the number of potential candidates to genes that are already connected to enhancer activity, such as the various co-factors discussed in Chapter 1. The maternal-Gal4 shRNA system also benefits from community resources such as the Transgenic RNAi Project, which has generated thousands of shRNA lines from which I can choose candidates to study[78].

The screen I describe in this chapter is also largely inspired by activity of the maternally-deposited transcription factor, Zelda. Zelda binds to essentially all known enhancers active in the early embryo. This binding precedes the actual function of these enhancers, suggesting that Zelda is a key activator of the zygotic genome and the presence of Zelda somehow primes enhancers for activation[65,79]. Zelda has also been shown to influence the chromatin environment by its association with the establishment or maintenance of open chromatin and deposition of the histone marks characteristic of active enhancers[40,41]. Like other transcription factors however, Zelda binds to many promoters in early development and is also essential for promoter function[65]. While I will not focus on Zelda specifically in this chapter, Zelda does serve as inspiration and its function implies that a system in which similar factors shape enhancer identity outside of the traditional model of transcription activation exists.

Others have also used Zelda as a proxy for discovery of Zelda-like factors before. Moshe and Kaplan (2017) characterized different chromatin properties around Zelda binding sites and used that to find similar features across the genome, revealing GAF also plays an important role in activating the zygotic genome[80]. Another group confirmed later confirmed this finding by showing that GAF is necessary for both zygotic genome activation and chromatin accessibility in the early *Drosophila melanogaster* embryo[81]. Using Zelda as a standard of comparison for novel factors, I will be able to screen for genes that affect enhancer activity and patterned gene expression.

In this chapter I show that knockdown of three different candidates by RNAi yields reduced embryonic viability, in addition to an altered patterned gene expression coupled with no effect of a ubiquitously expressed gene. Interestingly, both the viability and *in situ* hybridization results

vary depending on the particular fly line used for RNAi. My intent at the time was to follow-up on interesting candidates with single-embryo RNA-sequencing; however, I found unexpected variability of knockdown when examining *lid* expression. Following RNAi, *lid* expression levels of control embryos overlapped broadly with knockdown embryos using each shRNA line. Altogether the results show that a few candidates I have screened indeed may affect enhancer activity without affecting transcription in general; however, variability of gene expression upon knockdown precluded me from pursuing additional experiments using RNAi.

Methods

Fly Husbandry

All stocks were fed a molasses-based diet prepared by the UC Berkeley Media Prep Facility and maintained at 25 °C. I generated a nanos-Gal4; HisRFP line by crossing together Bloomington Drosophila Stock Center (subsequently referred to as BDSC) 4442 and 23650 as shown in Figure 2.1.

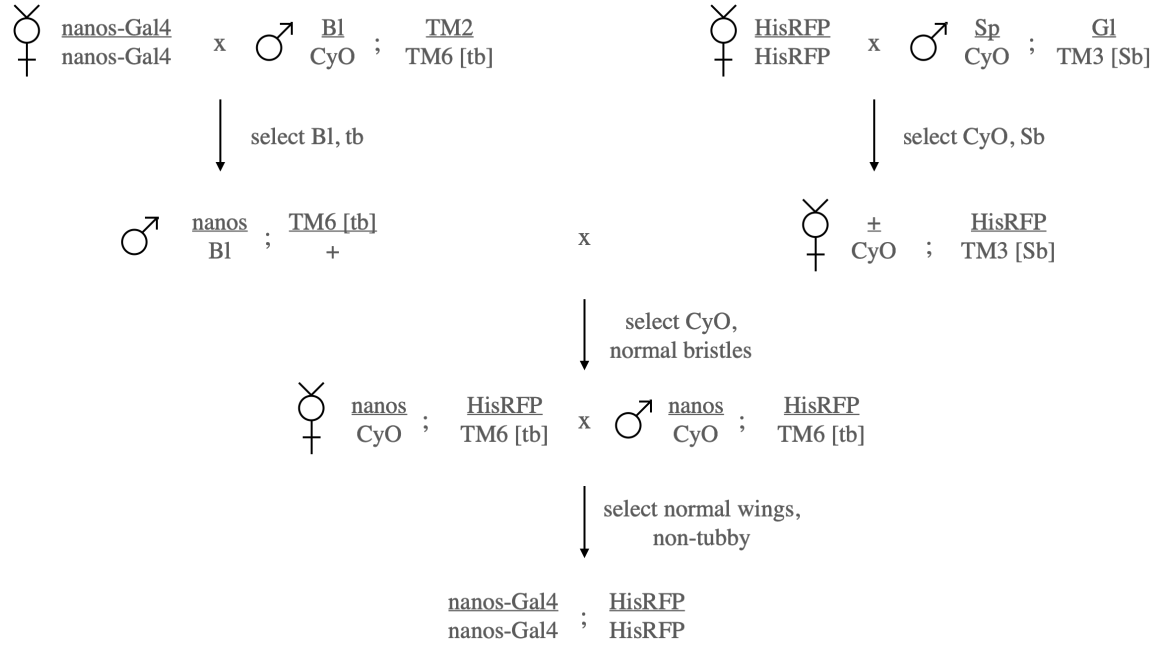

Figure 2.1 Crossing scheme used to generate nanos-Gal4; HisRFP lines In order to generate homozygous nanos-Gal4; HisRFP lines, I first had to separately balance nanos-Gal4 and HisRFP lines. I began the process of balancing nanos-Gal4 by crossing virgin females to double balancer males for chromosome 2 and 3 (Bl/CyO; TM2/TM6[tb]) as nanos-Gal4 is on chromosome 2, whereas HisRFP is on chromosome 3 so I needed to keep track of both chromosomes. At the same time, I began balancing HisRFP by crossing virgins to different double balancer males than the other cross in order to distinguish between the two in the next cross. For both crosses, I selected the desired genotypes by the markers indicated above. To generate double balanced flies containing both nanos-Gal4 and HisRFP I collected tubby CyO virgins with normal bristles and crossed them to sibling males with the same markers. After this cross, selecting flies that are wild-type for all markers and placing them into a new vial yields a homogenous stock.

I generated the matαtubulin-Gal4; HisRFP lines with the same crossing scheme. The HisRFP lines were generated with the intention of conducting follow-up experiments that would require precise staging of embryos. I did not conduct any experiments utilizing HisRFP in the end, nevertheless these lines were used for all RNAi experiments. All UAS-shRNA lines were obtained from BDSC and generated by the Transgenic RNAi Project[78]. I conducted the RNAi screen by crossing either nanos-Gal4; HisRFP or matalphatub-Gal4; HisRFP virgin females to UAS-shRNA males as indicated by the genes and Bloomington Drosophila Stock Center IDs shown in Table 2.1.

Gene	Reason for Interest	Bloomington Drosophila Stock Center ID(s)
Gug	co-repressor	51414, 32961
nej	chromatin modifier	36682, 37489
lid	chromatin modifier	36652, 35706
Su(var)3-3	chromatin modifier	36867, 32853
chameau	chromatin modifier	32484, 36869
Hdac3	chromatin modifier	34778
CG32264	intrinsically disordered protein	64966
Lilli	intrinsically disordered protein	34592
Alh	intrinsically disordered protien	39057
sba	intrinsically disordered protein	51488
CG10631	unidentified transcription factor	28001
CG12054	unidentified transcription factor	50511, 50910
CG12299	unidentified transcription factor	33957
CG32767	unidentified transcription factor	57720
CG9650	unidentified transcription factor	40852, 58323
CG12236	unidentified transcription factor	31949
CG5953	unidentified transcription factor	43257, 57543
CG8765	unidentified transcription factor	29447, 32343
Cp190	insulator	33903, 33944

CTCF	insulator	40850, 35354
Beaf-32	insulator	35642

Table 2.1 RNAi Screen Candidates Table contains information related to the genes screen (left), gene function (middle), and the corresponding Bloomington Drosophila Stock Center IDs (right).

Crosses were maintained at 25°C and placed into collection cages 3-4 days after hatching. Caged flies were fed with yeast paste made of Red Star yeast pellets and water and spread onto grape juice agar plates.

Viability Assay

To improve screening efficiency, I developed a viability assay to perform in lieu of examining gene expression patterns by *in situ* hybridizations straightaway. In order to both improve egg laying efficiency and avoid having to spread yeast paste on collection plates so I could see embryo and larvae better, I collected embryos on fermented grape juice agar plates as previously described[82] with a few modifications. I mixed one part frozen grape juice concentrate to one part water and sterilized the solution by boiling on a hot plate and aliquoted seventy-five mL of sterilized grape juice and 0.5 g Red Star yeast pellets into 250 mL Erlenmeyer flasks and placed these onto a large shaker at 30°C overnight. I added this mixture 1:1 with a 2% agar solution while still warm to the touch then poured into small petri dishes. I made a pre-determined amount of these plates fresh every experiment because the lack of anti-fungal, which if present would prevent fermentation, caused the plates to go bad within about a week even when stored at 4°C.

Because mothers will retain embryos that are aging for some period of time, I allowed them to lay on a fresh plate for 30 minutes to 1 hour to prevent collection of older embryos. I collected embryos on a new plate with minimal yeast paste placed in stripes for 1 hour, then hand-counted embryos under a dissection microscope by splitting the plates into quadrants and counting embryos within each quadrant. After counting, plates were returned to the 25°C incubator and covered. Thirty hours following the beginning of the collection, I placed the plates into a -20°C freezer to prevent larvae from crawling around the plate and affecting the final count. After 20-30 minutes I counted the larvae were counted in a similar manner as the embryos. Finally, I calculated viability by dividing the number of larvae divided by the number of embryos.

Fixation

I collected embryos for subsequent *in situ* hybridization on grape juice plates with minimal yeast paste and performing a 2 hour collections followed by 2 hours of aging to target nuclear cycle 14, the point at which the zygotic genome is largely activated. Embryos were transferred from plates into a collection basket made out of 50 mL conicals and mesh fashioned into the cap using a small paintbrush. The basket was placed and periodically agitated in 50% bleach for 3 minutes to remove the outer layer (chorion) of the embryo, then rinsed with cold water until the basket no longer smells like bleach. Embryos were then transferred into scintillation vials containing 3 mL of 1.3x Phosphate-buffered saline (PBS). Four mL of heptane and 1 mL 37% formaldehyde were added in the hood, then the vials were placed on their sides and taped to a shaker, shaken for 20

minutes at approximately 150 RPM. Embryos were then devitellinized by adding 10 mL methanol and vortexing for 30 seconds. After removal of the bottom phase, embryos were transferred into eppendorf tubes using a cut pipette tip, washed 3 times with 1 mL methanol and stored at -20 C until needed.

In situ hybridization probe synthesis

I generated *in situ* hybridization RNA probes for *eve* and *sryα* by amplifying the desired probe sequence from genomic DNA. Each reverse primer contained T7 promoter sequence added to the 3' end in order to incorporate either DNP-11-UTP or DIG-11-UTP into the final product. To amplify the *eve* sequence I used AA149 and AA150, and for *sryα* I used AA151 and 152 (see Table 2.2). For the *in* vitro transcription reaction and precipitation, I followed a modified protocol from Soile Keränen (June 2006). First, I followed the manufacturer's instruction for the T7 *in vitro* transcription reaction kit with the exception of my own 10x NTP mix containing 10 mM each of ATP, CTP, and GTP, with 6.5 mM of UTP and 3.5 mM of modified UTP and incubated the reaction for 4 hours at 37°C. To clean the final product I used Turbo DNase (ThermoFisher) according to manufacturer's instructions followed by a sodium acetate precipitation. I added 1/10th volume of 3M NaOAc pH 5.2 followed by 2.5 volumes of ethanol and placed the tubes into a -20°C freezer for at least 2 hours or until I could proceed with the next step. I spun down the precipitate in a 4°C centrifuge for 20 minutes at 13200 rpm. After removing the supernatant I washed the probes with 400 uL of 70% ethanol and spun for 20 minutes at 4°C and 13200 rpm. After removing the supernatant and briefly allowing the pellet to air dry, I added 25 uL of RNase-free water for every 10 uL of the *in vitro* transcription reaction. I quantified each probe on a nanodrop prior to adding 1 volume of formamide. Probes should be stored at -20°C.

Oligo	**Sequence**
AA149	TGAAGGACAAGCGTCAGAGG
AA150	caggtctgagtaatacgactcactataggCAATCACAGTTGTCGTCGGC
AA151	TTTGCCAATGTGGCCATTCATTC
AA152	caggtctgagtaatacgactcactataggCAGCAGCTCCTGGTTAAAATGCTCC

Table 2.2 PCR Primer sequences for generation of *in situ* hybridization probes Table contains sequences for oligos used to generate *eve* and *sryα* probes. Lowercase text corresponds to T7 promoter sequence.

Dual-color in situ hybridization

In order to perform dual-color *in situ* hybridization, I modified a previously published method developed by Soile Keränen[83] in combination with a standard colorimetric *in situ* protocol handed down to me by another lab member. I cleared the embryos by stepping them from methanol into ethanol, then rocking in a 90% xylene/10% ethanol for 1 hour. Following the clearing, I washed the embryos 3 times in EtOH, with the last wash nutating for 5 minutes. Then, I stepped the embryos into PBTween by nutating 4 times for 5 minutes each.

After stepping the embryos from PBTween into hybridization buffer (50% formamide, 5x SSC, 100 ug/mL salmon sperm DNA, 50 ug/mL heparin, 0.1% Tween-20), I prepared the embryos for

hybridization by incubating the embryos in 1 mL hybridization buffer for 1 hour in a 55°C waterbath. In the meantime, I pre-absorbed the anti-DIG or DNP-HRP antibodies in wild-type embryos to prevent any cross-reactivity. Following the pre-hybridization step, I boiled 2 uL of each probe in 200 uL hybridization buffer at 90°C for 3 minutes and snap-cooled on ice. I then removed the supernatant from the embryos, added the boiled probes and hybridized at 55°C overnight.

To remove the probes before adding antibodies, I performed a total of 6 washes in 55°C hybridization buffer over a period of 2.5 hours. I then stepped the embryos into PBT by nutating in 50% hybridization buffers and 50% PBTween for 15 minutes twice. Then I nutated the embryos in 1% BSA in PBTween 5 times for 10 minutes each.

After aspirating the supernatant, I added 200 uL anti-DNP HRP at 1:100 and nutated for 2 hours at room temperature covered in foil. I washed the antibody out with 3 quick washes in PBTween followed by nutating the embryos in PBTween for six 20 minute increments and an overnight wash in PBTween at 4°C. The next day I performed the first color reaction to detect *even-skipped* probe using a Tyramide Signal Amplification (coumarin) kit from Perkin Elmer. I stopped the reaction with 1 mL PBTween.

Prior to stripping the antibodies off of the embryos, I washed the embryos 5x quickly with PBTween. Then I stepped the embryos into hybridization buffer by washing in 50% PBTween/50% hybridization buffer for 5 minutes at 55°C. I washed the embryos 4x for 10-15 minutes in hybridization buffer at 55°C followed by 3X in PBTween for 15 minutes. To kill any remaining antibodies, I nutated the embryos in 5% formaldehyde in PBTween for 20 minutes. I stopped the reaction by washed the embryos quickly 3x in PBTween then nutating for 30 minutes in PBTween at room temperature. Occasionally I would leave the embryos nutating at 4°C if there was no time for further steps.

To detect the *sry*α probe, I added 200 uL of 1:100 anti-DIG HRP and incubated for 2 hours at room temperature covered in foil and followed the same washes as indicated above for anti-DNP HRP. I performed the second color rection the following day with another Tyramide Signal Amplification kit from Perkin Elmer in a different color (fluorescein or Cy5). I stopped the reaction in 1 mL PBTween followed by 5 quick washes in PBTween. I left the embryos in PBTween at 4°C for up to 48 hours prior to mounting the slides for imaging.

Mounting slides and confocal imaging

I made a 50 mL stock of mountant by mixing 1.25 g of DABCO (Sigma-Aldrich) crystals in 15 mL of 1x PBS and 35 mL glycerol into a light-shielded 50 mL conical and nutating until the solution is homogenous. For reference, DABCO mountant should be stored at -20°C.

I prepared the embryos for mounting by resuspending them in an arbitrary amount of DABCO mountant depending on how dilute I wanted the embryos to be on each slide. I allowed the embryos to settle for 1-3 hours at room temperature or 4°C overnight. Embryos can be stored for months or years in the mountant at -20°C as long as the embryos are shielded from light.

To mount the embryos, I transferred 35 uL of embryos onto a glass slide using a cut pipette tip between two cover slips placed roughly 7 mm apart. I then placed another cover slip on top of the two cover slips, this leaves enough space to not flatten the embryos during imaging. I sealed all of the edges of the coverslips with clear nail polish. After the nail polish completely dried I either proceeded with imaging or stored the slides at -20°C and imaged at a later date.

I imaged many embryos from each slide at various morphological stages using a Zeiss LSM 800 confocal and a 10x objective. For each embryo, I imaged 15 slices of half of the embryo using the 405 laser for eve-DNP + coumarin reaction and 488 laser for sryα-DIG + fluorescein reaction Due to the inconsistencies intrinsic to *in situ* hybridization and for these experiments the ability to detect any patterning changes was more important than using consistent laser power, the exact laser powers vary from embryo to embryo. I used Fiji/ImageJ to false colorize the embryos and generate maximum intensity projections to visualize the expression patterns.

Single-embryo RNA extraction and qRT-PCR

In order to quantify RNAs prior to zygotic genome activation, I collected embryos for 1.5 hours and aged for 70 minutes prior to extraction. I dechorionated the embryos by agitating in 50% bleach for 3 minutes. I collected six stage 3 embryos[84] by visualizing them under a dissection microscope and transferring the embryos with a paintbrush into six separate dounces containing 1 mL Trizol (Invitrogen). I isolated RNA as previously described[85]; however I slightly modified the homogenization of embryos in Trizol as indicated here. I homogenized the embryos in Trizol by douncing 10 times with a loose dounce (A) and tight dounce (B). At this point, I either kept the embryos frozen in Trizol at -80°C or continued with the manufacturer's instructions. After the RNA extraction was finished I further cleaned the RNA by using Turbo DNase (ThermoFisher) according to the manufacturer's instructions to remove any contaminating DNA.

I detected the levels of *lid* RNA in 8 control, and 8 knockdown embryos of each separate shRNA line using qRT-PCR. I amplified *lid* RNA with primer pair AA16 and AA17, and *Act5C* control RNA with AA12 and AA13 with the Invitrogen SuperScript™ III Platinum™ SYBR™ Green One-Step qRT-PCR Kit (see Table 2.3 below).

Oligo	Sequence
AA12	AGGCCAACCGTGAGAAGATG
AA13	ACATACATGGCGGGTGTGTT
AA16	TCGTCTGCGTACCCGAAAC
AA17	GGCTCGTCGTTAGCATTGGAT

Table 2.3 PCR Primers for one-step qRT-PCR Table contains sequences of primers used for qRT-PCR of *lid* and *Act5C* in the early *Drosophila* embryo.

I ran these samples on a Roche Lightcycler 480 at 50°C for 3 minutes (cDNA synthesis), 95°C for 5 minutes, followed by 40 cycles of 95°C for 15 seconds, 55°C for 30 seconds, and 72°C for

60 seconds. I then held the reaction at 40°C for 1 minute followed by melting curve analysis to validate the primers.

Using R, I normalized *lid* expression to *Act5c* expression by using each Cp value as input and exponentiating the value obtained from a previously generated standard curve per primer pair in order to get relative expression of each gene from each embryo, then divided *lid* expression by *Act5c* expression to obtain a single value per replicate (code accompanying related preprint is available here: https://github.com/aralbright/2021_AAME_qPCR). I generated qRT-PCR figures using ggplot2[86].

Results

Knockdown of transcriptional co-factors results in reduced viability

Knowing that many transcription factors may bind to co-factors or affect chromatin remodeling (reviewed in [37]). I hypothesized that these co-factors and chromatin remodelers are involved in specifying enhancer identity and activity. I also examined various factors of general interest to the lab, namely intrinsically disordered proteins and insulators, as well as unidentified transcription factors (see Table 2.1). Because most of the early patterned genes and transcription factors are essential for viability, and the loss of Zelda specifically becomes lethal prior to the larval stage, I decided to design a viability assay to determine how many knockdown embryos survive embryogenesis. When conducting this assay and subsequent experiments, I used two driver lines because each is active at a different stage of oogenesis and may have different effects on embryonic development and two RNAi lines (if available) per candidate because of variable knockdown efficiency.

Gene	BDSC	Driver	Replicate	Embryos	Larvae	Viability
Gug	32961	nanos	1	250	0	0.0
		nanos	2	160	0	0.0
		matα	1	843	0	0.0
		matα	2	142	0	0.0
	51414*	nanos	1	232	244	1.05
		nanos	2	437	402	0.92
		matα	1	55	55	1
		matα	2	33	27	0.82
nej	37489	nanos	1	38	1	0.03
		nanos	2	27	1	0.04
		matα	1	309	189	0.61
		matα	2	49	11	0.22
	36682	nanos	1	594	155	0.26
		nanos	2	299	9	0.03
		matα	1	496	24	0.05
		matα	2	260	162	0.62

lid	35706	nanos	1	544	149	0.27
		nanos	2	357	38	0.1
		matα	1	606	437	0.72
		matα	2	228	74	0.32
	36652	nanos	1	587	248	0.42
		nanos	2	232	79	0.34
		matα	1	295	276	0.93
		matα	2	262	137	0.52
Control	nanos	nanos	1	532	410	0.77
		nanos	2	212	130	0.61
	matα*	matα	1	10	9	0.9
		matα	2	14	13	0.92

Table 2.4 Viability assay to determine if candidate genes are necessary for enhancer function Table depicts results from a single experiment where all crosses and collections were done simultaneously to eliminate environmental effects, with the exception of results depicted with a * which were collected on a different day. The matα flies in particular were unhealthy and never laid enough embryos. After collecting embryos from each respective cage for an hour, I counted the total number embryos on the plate. Thirty hours later, I counted the number of larvae on the plate. I calculated viability as the number of larvae divided by the number of embryos. Although I placed plates in the freezer to hopefully slow down the larvae, the larvae still moved and likely contribute to human counting error as indicated by a viability of greater than 1 in the table above.

I designed this viability assay for counting larger numbers of embryos and larvae, and to observe large changes of viability in knockdown embryos relative to the controls. As such, this viability assay is not designed to obtain exact calculations of viability, as I obtained a viability value of 1.05, which is not physically possible. Using Zelda as an example of what I should be looking for in this screen, loss of candidate gene should result in lethality, or 0% viability. As a matter of fact, one positive control (*Gug*) almost phenocopies Zelda with one RNAi line resulting in 0% viability. Overall, I found multiple candidates that yield greatly reduced viability upon knockdown (see Table 2.4). I conducted this viability assay for all candidates listed in Table 2.1, but I decided to focus my attention on *lid, Gug,* and *nej* because I did not see any reduction in viability for any of the other genes and decided not to pursue them any further.

Knockdown of transcriptional co-factors results in loss or alteration patterned gene expression while ubiquitous expression is unchanged

If a candidate gene knockdown does have an effect on enhancer identity and activity specifically, in theory embryos would lose patterned gene expression while ubiquitous or housekeeping gene expression remains the same. Following the viability assay, I performed dual-color *in situ* hybridizations on fixed 2-4 hour knockdown embryos for a patterned and non-patterned gene. Because I am limited to doing two *in situ* hybridizations per experiment, one gene must undoubtedly represent enhancer-patterned transcription and the other represent transcription in general.

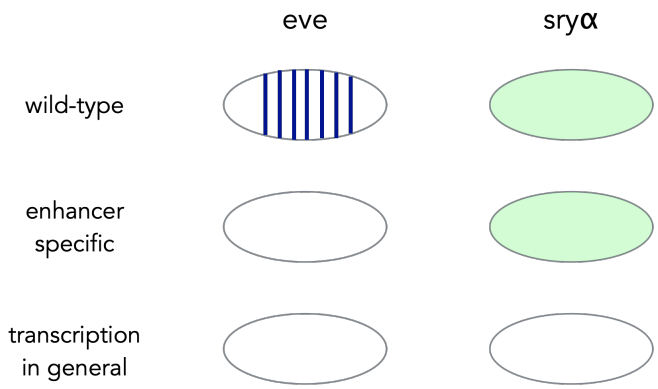

Figure 2.2 Diagram of RNAi screen outcomes Figure shows cartoon representations of three possible outcomes of *eve* patterned expression and *sryα* ubiquitous expression following RNAi. RNAi with no effect on enhancer activity or transcription in general will yield wild-type patterning, enhancer-specific effects will alter *eve* patterning while *sryα* shows normal expression, transcription defects in general will result in changes in expression of both genes. Results depicted are not necessarily the only outcomes; however, this diagram frames the organization of my thoughts throughout conducting these experiments.

Eve is one of the most highly studied patterned genes with a clear refined pattern at nuclear cycle 14, and *sryα* is expressed ubiquitously and reliably at the same time point. In this screen, *eve* serves as a proxy for patterned gene expression and *sryα* as a proxy for non-patterned expression, or transcription in general. Knockdown of the ideal candidate would alter or eliminate *eve* expression, while *sryα* expression remains normal. A scenario where both *eve* and *sryα* expression are affected would be interesting, but the candidate gene would not meet the enhancer-specific ideal (See Figure 2.2).

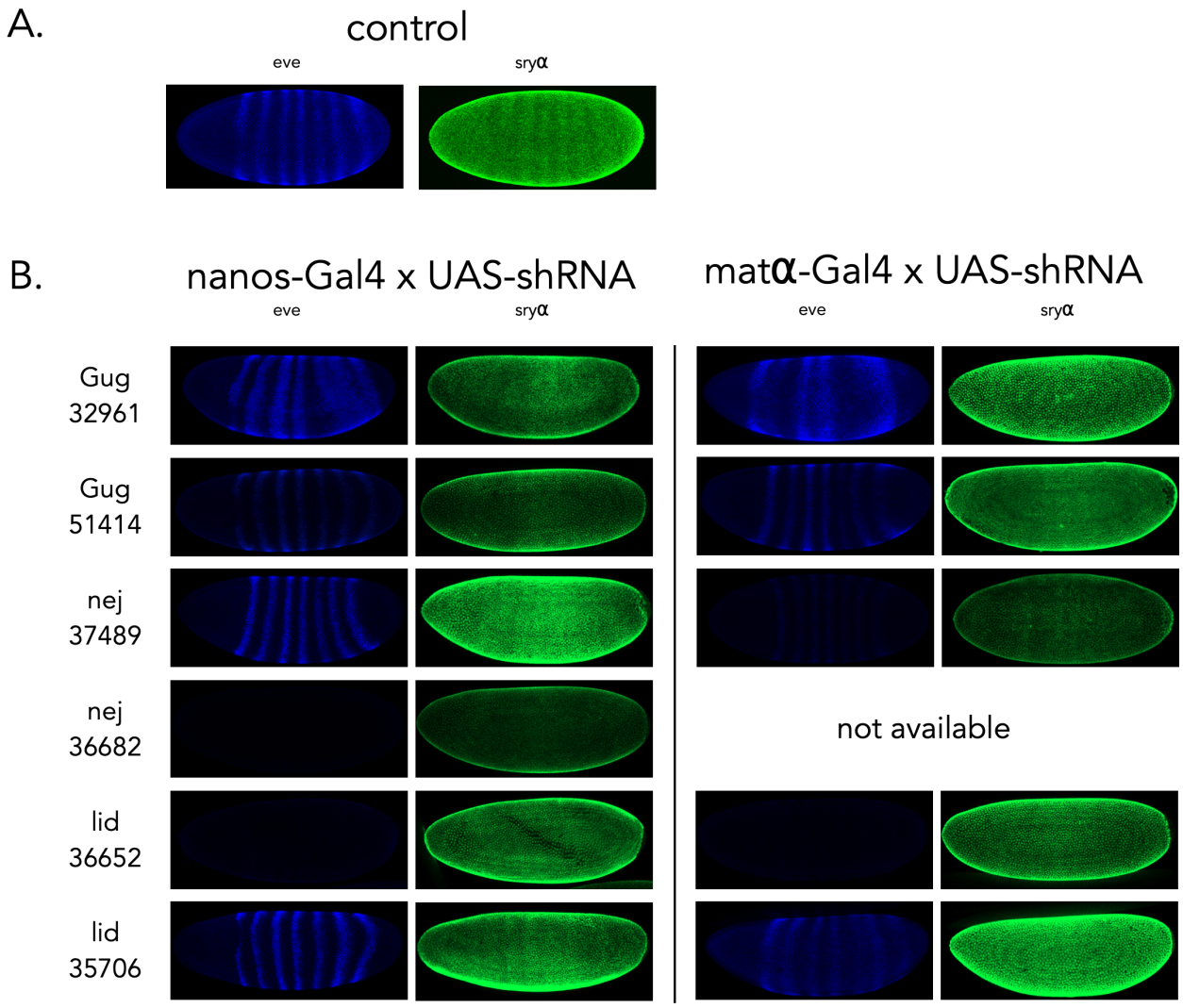

Figure 2.3 RNAi knockdown results in altered patterned gene expression with no apparent effect on ubiquitous gene expression depending on the driver and RNAi line used Figure depicts representative images of dual-color *in situ* hybridizations of *eve* (blue) and *sryα* (green) in (A) control and (B) RNAi knockdowns of candidates involved in the establishment of enhancer identity and/or activity. I was unable to obtain images of *nej* 36682 embryos because these embryos were unhealthy.

While I found that knockdown of *Gug* using line 32961 was entirely lethal from the viability assay (see Table 2.4), the embryos still express *eve* but in an abnormal pattern with normal *sryα* expression (see Figure 2.3). An *eve* patterning defect upon loss of *Gug* is known from prior work[87], but this functions as a useful positive control indicating that the methods work and can produce interesting results.

I should note that both RNAi and *in situ* hybridizations can produce variable results. The unaffected viability of *Gug* 51414 knockdown embryos combined with no apparent effect of *eve* patterned expression highlight the importance of using multiple driver and shRNA lines. The same case can be made for the *nej* knockdown embryos where both lines may show some indication of reduced viability, but *eve* patterning is retained in some cases. *Nej* encodes a well

described essential co-activator, with early lethality upon loss[88]. The variable phenotypes presented here are inherently interesting, but the fact that these two lines are positive controls in a sense because we either know the phenotype already (*Gug*) or know the importance of already (*nej*), I decided to further pursue *lid* as an interesting candidate.

Knockdown of *lid* appears to result in reduced viability and complete loss of *eve* expression. Viability following *lid* knockdown does vary (see Table 2.4); however, knockdown with both Gal4 drivers consistently yields similar results with a complete loss of *eve* expression. This is unlikely to be an experimental artifact, as I always processed many samples at the same time using the same probe and reagents. With that being said, the phenotype is not consistent between the two RNAi lines as knockdown with 35706 does not completely remove *eve* expression; however, the viability of embryos using both RNAi lines is greatly reduced upon knockdown. The results of the *in situ* hybridizations are inconsistent, but this could be due to inefficient knockdown by RNAi. I proceeded by using qRT-PCR to quantify the efficiency of knockdown.

Inconsistent knockdown of *lid* expression revealed by single-embryo qRT-PCR

Previously published work showed that in bulk, 35706 and 36652 RNAi lines using the matαtubulin-Gal4 driver efficiently knocked down *lid* expression in mature oocytes[89]. I was unable to repeat this experiment in embryos as my matαtubulin-Gal4 stocks became sick and yielded flew flies after crossing, but I was able to move forward with confirming knockdown using the nanos-Gal4 driver. My intentions at the time were to characterize gene expression further using single-embryo RNA-sequencing and I felt that demonstrating effective RNAi in bulk was not sufficient to proceed. At this point, I decided to assess the levels of knockdown in single embryos, which to my knowledge for the purpose of validating RNAi had never been done before.

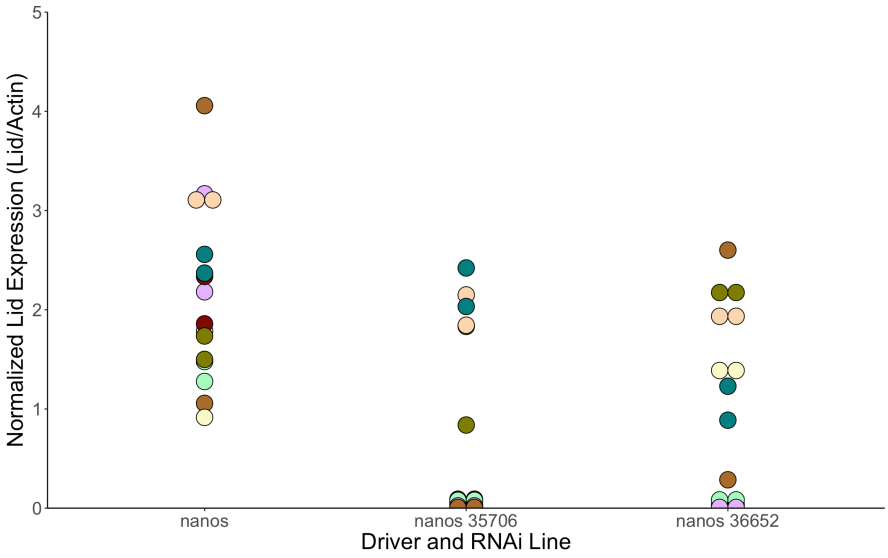

Figure 2.4 Wide range of normalized *lid* expression in individual control and knockdown embryos. Figure shows relative levels of *lid* normalized using *Act5c* (Actin) expression in Stage 3 embryos. For

each condition (control, RNAi line 1, RNAi line 2) I extracted RNA from 10 individual embryos and carried out qRT-PCR for each line with two replicates each. Each color represents an individual embryo from the corresponding condition.

Surprisingly, I found that normalized *lid* expression varied widely both within and between control and knockdown conditions. With such a wide range of overlap of *lid* expression for both wild-type and knockdown embryos shown in Figure 2.4, understanding the differences in gene expression with single-embryo RNA-sequencing due to natural variation versus a knockdown effect would be difficult. Technical replicates for the most part, appear to coincide with one another; however, as an example the two replicates of the brown sample in nanos 36652 are not close. With that being said, clear biological variation exists within these data as well given the range of expression between different samples. At this point, I decided the variability of knockdown was too great to continue with this approach to understanding the regulation of gene expression by enhancer activity upon loss of candidate factors. Ultimately, *Gug, lid,* and *nej,* yielded interesting phenotypes from this screen; however, additional experiments must be conducted in order to understand their role in the establishment of enhancer activity and identity in the early *Drosophila melanogaster* embryo.

Discussion

Understanding the specification of enhancer activity and identity is essential to our understand of early development. In *Drosophila melanogaster* embryos, the zygotic genome is activated in the 14th nuclear cycle. Until then, development is largely controlled by maternally-deposited RNAs and proteins, thus maternal factors must influence enhancer activity and identity. Here, I used the maternal-Gal4 shRNA system to knockdown genes, such as co-factors and chromatin modifiers, that I hypothesized to be important in establishing enhancer-specific activity. From this study I found three interesting candidates (*lid, Gug,* and *nej),* all chromatin modifiers, that show reduced viability upon knockdown by RNAi. Additionally, *eve* expression is either lost or altered in some way while *sryα* expression remained normal upon knockdown of each of these candidates.

Considering *eve* as a proxy for enhancer-patterned expression and *sryα* expression as a proxy for transcription in general, these results that each candidate may play a role in establishing enhancer identity outside of their normal role as chromatin modifiers. With RNAi, I cannot rule out other candidates screened as knockdown may have been inefficient in cases where I saw no effect on viability. With that being said, eliminating other candidates listed in Table 2.4 from further study because I did not observe a phenotype does not necessarily mean that these genes are not involved in the specification of enhancer identity. RNAi knockdown can be variable for a number of reasons, but I chose to focus on genes where I noticed a clear reduction in viability.

Using *in situ* hybridizations, I was limited to examining expression patterns of only two genes at a time. At the beginning of my PhD, no method existed to understand patterning in many genes across the genome, therefore I intended to conduct single-embryo RNA-sequencing to at least examine expression level of all genes in a single individual. This would not provide any information on patterning, but examining patterning in select genes with a large change in expression level remained a possibility.

As shown in Figure 2.4 however, confirming knockdown by single-embryo qRT-PCR prior to sequencing revealed unexpected variability in the degree of knockdown by RNAi. Considering that expression level of *lid* was widely variable upon knockdown, I questioned whether the expression of many other genes would also be too variable to draw reasonable conclusions from in single-embryo RNA-sequencing. I felt that sharing the variability of knockdown by RNAi was important, thus I published part of this work as a preprint where I further discuss how the analysis may effect these results[73]. At this point, I decided to pursue interesting candidates using germline clones instead. Fortunately, this decision coincided with increasing availability of single-cell and nucleus RNA-sequencing technologies that would allow me to examine spatially resolved gene expression, which I will discuss in the following chapter.

Chapter 3: Demonstrating the use of single-nucleus RNA-sequencing to examine patterned gene expression in the early *Drosophila* embryo

Abstract

Our current understanding of patterned gene expression comes either from analyses observing patterning in a few genes at a time, as with *in situ* hybridizations, or observing levels of expression without regards to patterns, as with RNA-sequencing. In order to understand the regulation of many genes in the early *Drosophila* embryo at single-nucleus resolution, I have conducted single-nucleus RNA-sequencing in pre-cellularization embryos upon loss of maternally-deposited dCTCF. I demonstrate that after adequate filtering of the sequencing data, which are noisy, the nuclei cluster on patterned gene expression according to their original spatial position prior to nuclear isolation. Additionally, I have shown that this technique has the potential to find genes that are differentially expressed in one or more clusters, but not in bulk. This would suggest that gene expression is differentially affected in different regions of the embryo. Ultimately, our ability to conduct single-nucleus RNA-sequencing in the early *Drosophila* embryo will to allow us to ask questions regarding the regulation and establishment of patterned gene expression at a larger scale than we previously could.

Introduction

The early *Drosophila melanogaster* embryo has long been a system used for studying the regulation of gene expression, particularly of patterned genes, in early development. Many features associated with enhancer activity, as discussed in Chapter 1, affect the expression of patterned genes; however, insulators are of particular interest because less is known about their role in patterned expression. Insulators are elements of the genome that drive chromatin organization by creating stable physical separation between adjacent domains[50]. Several studies show that the loss of an insulator protein, dCTCF, impacts patterned gene expression at a few genes[54,55]. However, other studies suggest that loss of insulator proteins in the early embryo has minimal impact on gene expression globally[46,52,59,60]. Given the importance of dCTCF in maintaining genome structure and expression of certain patterning genes, *dCTCF* is a good candidate to use in establishing the use of single-nucleus RNA-sequencing in the early *Drosophila* embryo.

Early *Drosophila melanogaster* development begins with 14 rapid nuclear divisions in a syncytium, or a group of nuclei not separated by a plasma membrane. Maternally-deposited RNAs and proteins drive early development and establishing the regulation of gene expression prior to zygotic genome activation (reviewed in [63]). With that considered, we must be able to distinguish between maternal and zygotic RNAs experimentally when asking questions about zygotic gene expression. This can be accomplished by crossing genetically distinct lines and detecting polymorphisms in the obtained sequences following single-embryo RNA-sequencing[85]; however, this would not be feasible when generating candidate mutants for analysis. Single-cell RNA-sequencing after gastrulation is a possibility, and has been demonstrated[71,72]; however,

zygotic genome activation occurs prior to cellularization. In order to understand the regulation of gene expression at a large scale concurrent with zygotic genome activation, development of methods to observe gene expression in nuclei are necessary.

Isolating nuclei and conducting single-nucleus RNA sequencing would allow for the study of patterned gene expression without contamination of cytoplasmic, or maternally deposited RNAs, on a larger scale than ever before. With or without the ability to recapitulate expression patterns from single-nucleus data computationally, we would be able to examine gene expression in different groups of nuclei upon some perturbation. This would allow us to answer questions regarding the potential for genes to be differentially expressed in different nuclei, which can be confirmed with further experiments.

In this chapter, I demonstrate the use of single-nucleus RNA-sequencing in early *Drosophila melanogaster* embryos to examine changes in zygotic gene expression upon loss of maternal dCTCF. I found that genes marking each cluster of nuclei are expressed distinctly relative to other clusters, and the marker genes representing most of the clusters correspond to spatial regions of the embryo. I also found many examples of genes differentially expressed in one or more clusters, but not in bulk. The distinct expression and differential expression of genes that I found within and between clusters highlights the potential for single-nucleus RNA-sequencing to answer questions surrounding the regulation of gene expression across different regions of the embryo.

Methods

Fly husbandry

All stocks were fed standard Bloomington food from LabExpress and maintained at room temperature unless otherwise noted. I used the FLP-DFS (dominant female sterile) technique[77] to generate germline clones, or maternal nulls, of *dCTCF*. First, I created a *hsFLP; Gl*/TM3(Sb)* stock by combining Bloomington Drosophila Stock Center #8862 (hsFLP) and a *Gl*/TM3(Sb)* balancer stock as shown in Figure 3.1. Then, my labmate Michael Stadler generated a recombinant *dCTCF*, FRT2A* line as shown in Figure 3.2. Finally to generate germline clones, I conducted the cross as shown in Figure 3.3.

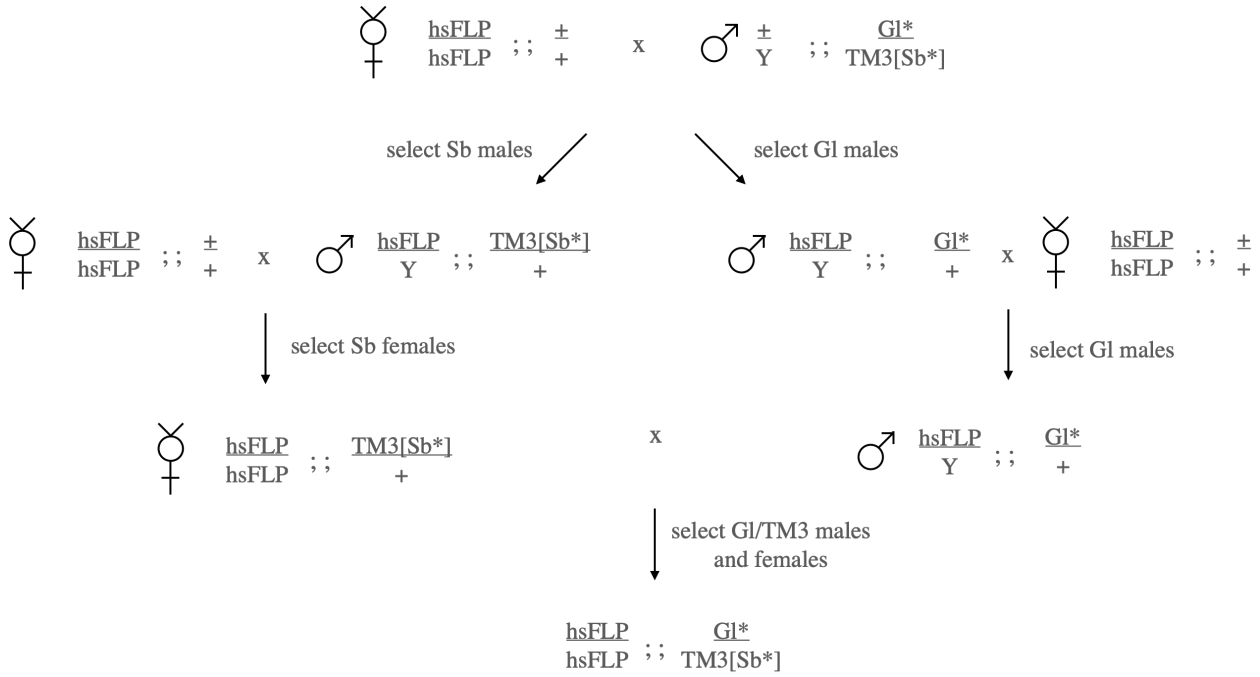

Figure 3.1 Crossing scheme to generate hsFLP line Figure outlines the steps taken to generate a third chromosome balanced hsFLP line.

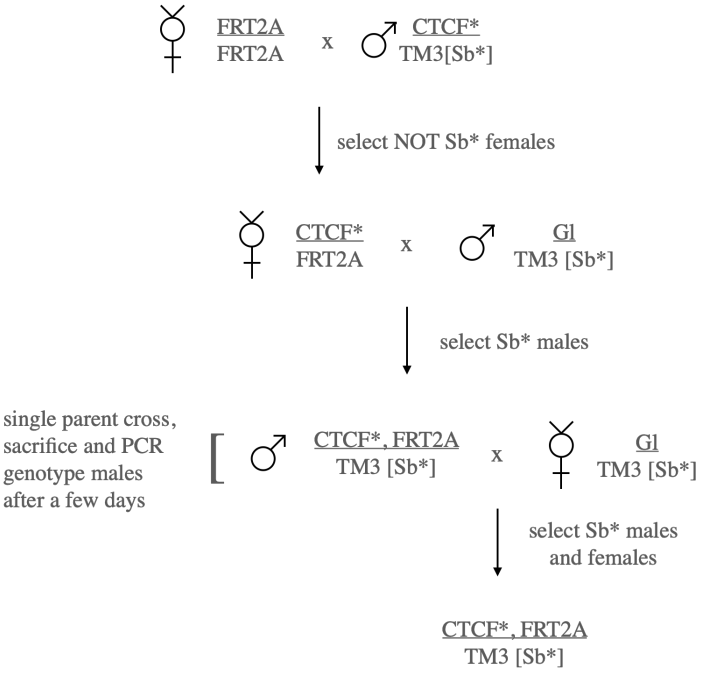

Figure 3.2 Crossing scheme to generate CTCF, FRT line Figure outlines the steps taken to generate a recombinant CTCF*, FRT2A line balanced on the third chromosome. This CTCF* mutant and cross as described above were generated by my labmate, Michael Stadler.

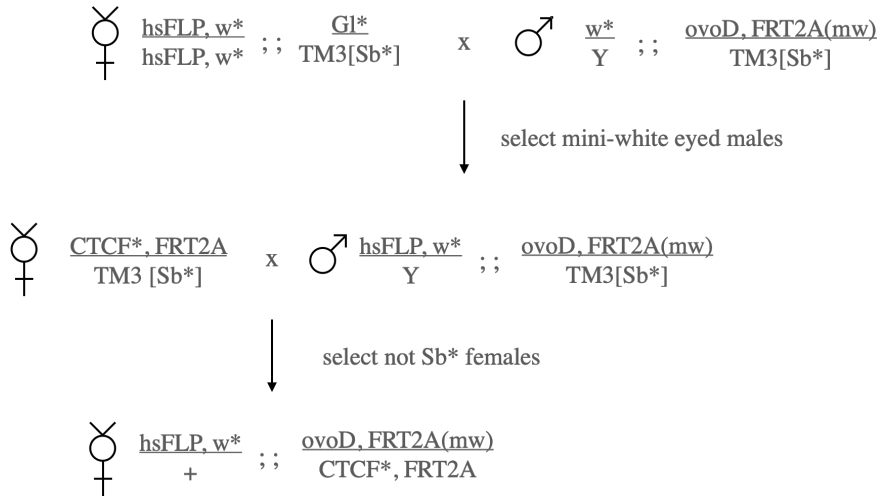

Figure 3.3 Crossing scheme to generate germline clones Figure depicts cross to generate maternal *dCTCF* null embryos.

After the last cross depicted in Figure 3.3, I heat-shocked the vials for 2 hours on days 4, 5, and 6 following the cross to induce recombination. When these flies hatched, I placed non-stubble female flies and their male siblings into a medium cage and fed them every day with Red Star yeast paste spread onto an apple juice agar plate.

Nuclear Isolation

I isolated nuclei from embryos according to a protocol based on a combination of previously published work[32,90,91]. I collected Stage 5 (nuclear cycle 14) embryos from each cage by clearing the cages for 30 minutes to 1 hour, followed by a 2 hour collection at 2 hour aging. Prior to sorting the embryos, I dechorionated the embryos in 100% bleach for 1 minute with agitation directly on the molasses plate until the embryos were floating. I transferred the embryos to a collection basket made out of a 50 mL conical and mesh. I rinsed the embryos in water before transferring the embryos to a 1.5 mL eppendorf tube containing 0.5% PBST. At this point, I kept everything on ice to prevent any further aging of embryos and nuclei.

I selected a minimum of 9 early to mid NC14 embryos using an inverted compound light microscope and transferred them to a 2 mL dounce containing 600 uL of lysis buffer (10 mM 10 mM Tris-HCl pH 7.4, 10 mM NaCl, 3 mM MgCl2, 1% BSA, 1% RNase Inhibitor) + 0.1% IGEPAL. I homogenized the embryos 20 times with a loose pestle and 10 times with a tight pestle, rinsing the pestles with 100 uL lysis buffer + 0.1% IGEPAL after completion in order to reduce loss. I transferred 800 uL of buffer and embryos to an eppendorf tube and filtered using a 40 uM filter.

I pelleted the nuclei by spinning for 5 minutes at 900 g at 4°C. I washed the nuclei in 500 uL lysis buffer without IGEPAL and pelleted again before resuspending the nuclei in 20 uL lysis buffer without IGEPAL.

UC Berkeley Facilities Library Preparation and Sequencing

I would like to thank Dr. Justin Choi from the UC Berkeley Functional Genomics Laboratory for running the 10x Chromium Controller with 3' Reagent Kit (v3). Dr. Choi also ensured that the nuclei concentration was approximately 1000 nuclei per uL prior to running each sample. I would also like to thank the UC Berkeley Vincent Coates Genomics Sequencing Laboratory for sequencing these samples on the Illumina NovaSeq.

Packages Used and Code Availability

I analyzed the data in Python and R, using primarily scVI[92], scanpy[93] and custom scripts for analysis. All code used to analyze the data and generate figures below is available here: https://github.com/aralbright/2021_AAMSME

Data Preprocessing

To generate a nucleus x gene matrix containing UMI counts for each gene from raw sequencing reads, I first generated a custom reference index and then continued processing control and *dCTCF*$^{mat-/-}$ experiments separately according to the kallisto-bustools workflow[94]. I used custom python scripts to visualize quality control metrics. I also used custom scripts along with scanpy[93], a python package to analyze single-cell or nucleus data: (1) filter the nuclei according to the expected number of nuclei (10,000), (2) remove nuclei with fewer than 200 expressed genes, (3) remove nuclei with greater than 5% mitochondrial expression, (4) remove nuclei with greater than 50,000 UMI counts, and (5) remove genes expressed in fewer than 3 nuclei.

Batch Correction

Prior to batch correction, I found the top 6000 highly variable genes on log1p normalized data (normalized to counts per 10,000) using sc.pp.highly_variable_genes and subset the nucleus x gene matrix to those genes in order to reduce noise in downstream analyses. I then ran scVI on the raw (not normalized or transformed in any way) data with gene_likelihood set to "nb."

I obtained corrected values of gene expression for downstream analyses using scVI's model.get_latent_representation and normalized expression using model.get_normalized_expression, with library_size set to 1e4.

Clustering

I used the Leiden[95] clustering algorithm and scanpy[93] to cluster the nuclei and generate a 2D UMAP[96] for visualization. Before batch correction, I clustered the data on log1p normalized expression. After batch correction, I clustered the data on the latent space derived from the scVI model. Gene expression per nucleus on the UMAP plots represent the log of normalized scVI-derived expression.

Marker Gene Analysis

To find marker genes that represent each cluster, I used scanpy's sc.tl.rank_genes_groups function using the Wilcoxon signed-rank test[97] and sc.pl.rank_genes_groups_heatmap to visualize scaled gene expression of each marker gene within each nucleus stratified by cluster.

I determined which region of the embryo each cluster belonged to using *in situ* hybridizations (conducted by the Berkeley Drosophila Genome Project[98–100]) of the top marker genes.

Differential Expression

Using diffxpy (https://diffxpy.readthedocs.io/), I obtained log 2 fold change (log2FC) and associated p-values between control and *dCTCF* $^{mat-/-}$ nuclei in bulk and within each cluster. Then, I Bonferroni adjusted the p-values and filtered for significance with adjusted p-values < 0.05 and the absolute value of log2FC <= 1.5.

Before determining which clusters share sets of differentially expressed genes, I corrected the adjusted p-values by multiplying by the total number of conditions compared (adjusted p-value * 11). I then generated the UpSet plot to find intersecting genes between sets of clusters or bulk data using UpSetR[101,102].

Results

Single-nucleus RNA-sequencing data are noisy and must be filtered

The initial nucleus x gene matrix contains over 200,000 barcodes (or nuclei) for both control and *dCTCF*$^{mat-/-}$ experiments. Considering that a single 10x Genomics run captures a maximum of 10,000 nuclei, and that many droplets are empty, these data must be filtered prior to any meaningful analysis.

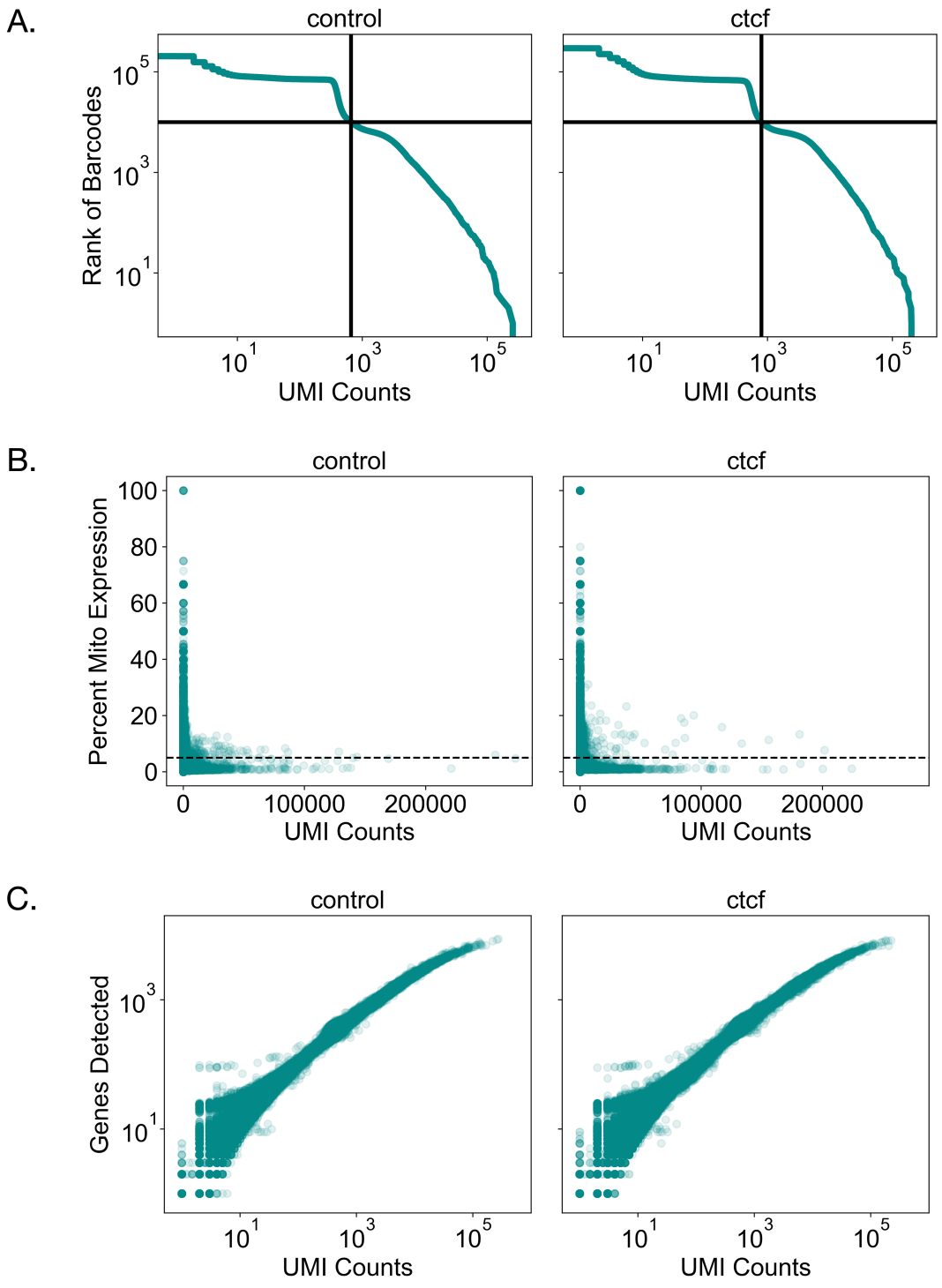

Figure 3.1 Quality control before filtering Figure depicts three different quality control metrics for control (left) or *dCTCF* $^{mat-/-}$ (right) nuclei. (A) Knee plot of barcodes ranked by the total UMI counts versus the number of UMI counts for control (left) and *dCTCF* $^{mat-/-}$ (right). The solid black lines indicate the position of the 10,000th nucleus on the x and y axis. (B) Percent mitochondrial expression by UMI counts for control (left) and *dCTCF* $^{mat-/-}$ (right). The dashed black lines indicate the 5% percent mitochondrial expression filter used downstream. Any nucleus above the dashed line will be removed from the dataset. (C) Library saturation plots, or the number of genes detected according to the UMI counts.

First, in order to distinguish between an empty droplet and one which contains a nucleus, I generated knee plots (Figure 3.1.A). The knee plot, first described for this purpose by Macosko et al. 2015[103], is used to observe the transition between droplet containing a nucleus and an empty droplet. This is represented by the inflection point, or the point at which the derivative of the curve is minimized. The maximum number of expected nuclei in each experiment is 10,000. By indicating this point with black lines coming from its position on the x and y-axis, I can see by eye that the expected number of nuclei corresponds closely to the inflection point. Any nuclei to the left of this point in Figure 3.1.A will be removed from the dataset.

Percent mitochondrial expression is another commonly used metric for filtering single-nucleus (or cell) RNA-seq data. The best cutoffs for different systems have been systematically determined, and different cell types might contain different amounts of mitochondrial RNAs depending on their function[104]. However, I was not sure what to expect with early *Drosophila* embryonic nuclei. In both the control and *dCTCF*$^{mat-/-}$ experiments, values below 5% seemed reasonable given the spread of values under 5% across UMI counts (Figure 3.1.B). Any nucleus with greater than 5% mitochondrial expression was removed from the dataset.

To determine if I had sufficiently sequenced enough of each library, I examined the number of genes detected according to the total number of UMI counts per nucleus. In theory, sequencing more (or a higher UMI count by proxy) should lead to the detection of more genes; however, sequencing is considered saturated once the number of genes detected stops increasing. As shown in Figure 3.1.C for both control (left) and *dCTCF*$^{mat-/-}$ (right), I could probably detect more genes if I sequenced more, but the rate at which new genes are detected tapers off. At this point, I decided I did not need to sequence any further.

In the end, I filtered the data according to the top 10,000 nuclei ranked by barcode, then for nuclei with less than 5% mitochondrial expression. I also decided to remove nuclei that expressed under 200 genes and remove genes expressed in less than three nuclei because these are likely not informative. Occasionally, a barcode might represent more than one nucleus so I removed nuclei with UMI counts greater than 50,000. In order to determine if these filtering metrics are adequate, I conducted several additional analyses prior to asking questions regarding gene expression in each nucleus.

Raw and normalized control gene counts are correlated with *dCTCF*$^{mat-/-}$ gene counts

In order to ensure that any downstream differences in gene expression are biological and not simply a result of having vastly different datasets, I calculated the average raw expression of each gene across all nuclei in the control and *dCTCF*$^{mat-/-}$ datasets as well as the average normalized expression, where values within each nucleus were normalized to counts per 10,000.

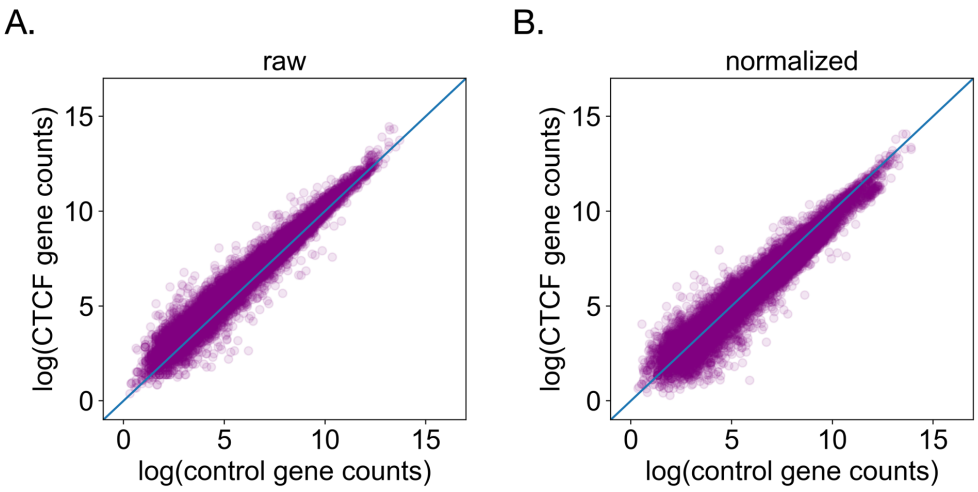

Figure 3.2 Average gene counts appear correlated between control and *dCTCF*$^{mat-/-}$ samples Figure depicts log(mean) UMI counts, where each point is a gene, for (A) raw and (B) normalized UMI counts. The blue line in both panels corresponds to y = x.

Clearly for both raw (Figure 3.2.A) and normalized (Figure 3.2.B) UMI counts, the vast majority of the lie close to the y = x line, indicating high correlation between the control and *dCTCF*$^{mat-/-}$ datasets. At this point, I decided to move forward with clustering the nuclei.

Batch correction is necessary to eliminate non-biological variation

Although the two datasets are highly correlated as shown in Figure 3.2, processing samples on different days, and using different 10x chips among other things, can lead to non-biological variability within the data, also referred to as batch effects. Because *dCTCF*$^{mat-/-}$ embryos survive embryogenesis to some extent[54], and mean gene expression is correlated between control *dCTCF*$^{mat-/-}$ embryos (see Figure 3.2), I would expect the nuclei to overlap in reduced dimensional space after clustering.

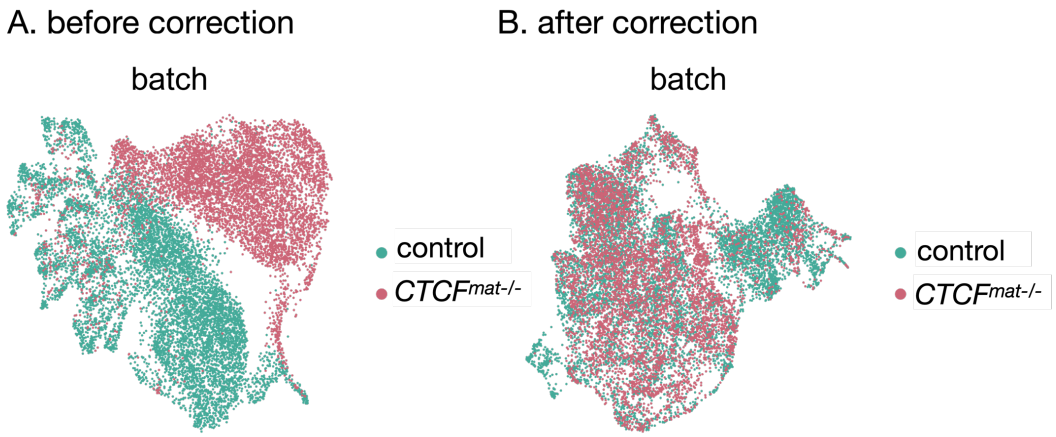

Figure 3.3 Batch correction integrates control and *dCTCF*$^{mat-/-}$ datasets Figure represents a 2D UMAP projection of control (teal) and *dCTCF*$^{mat-/-}$ (pink) nuclei (A) before and (B) after batch correction.

However, this is not the case prior to batch correction (see Figure 3.3.A). After removing background variability with scVI through scVI-tools[92,105] (see Figure 3.3B), the nuclei are better integrated as expected. Once the data were corrected, I wanted to ensure that the data were filtered to the best of my ability with further quality control.

Clustering on quality control metrics and expression of undesirable genes warrants cluster removal

Before proceeding with the analysis, I decided to evaluate the quality of the clusters by first seeing if the nuclei cluster on any of the previously mentioned quality control metrics. It is possible that different types of nuclei express more genes, have higher amounts of RNA, or mitochondrial RNA specifically. However, I decided this was less likely than previous filters not eliminating poor quality nuclei.

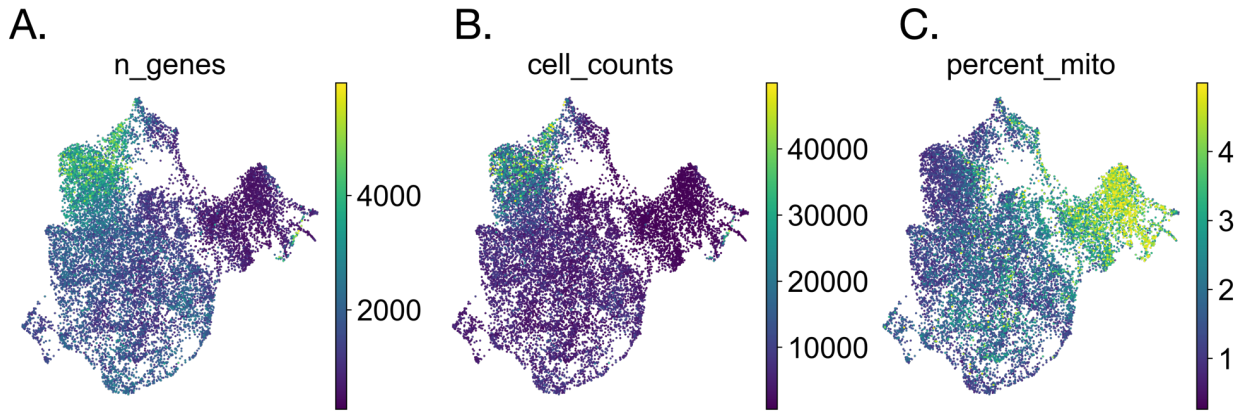

Figure 3.4 Nuclei appear to cluster on quality control metrics Figure depicts 2D UMAP representation of all nuclei colored by the (A) number of genes expressed in each nuclei, (B) the total UMI counts per nucleus, and (C) percent mitochondrial expression in each nucleus.

For adequately filtered data, I would expect for nuclei to only cluster according to biological variation in the data. Although it is possible for certain nuclei to express higher number of genes or percent mitochondrial expression, I ultimately decided to take a conservative approach by removing clusters that appeared to cluster on the quality control metrics as shown in Figure 3.4. Before re-clustering the data however, I decided to first determine if the nuclei cluster on expression of known nuclear markers.

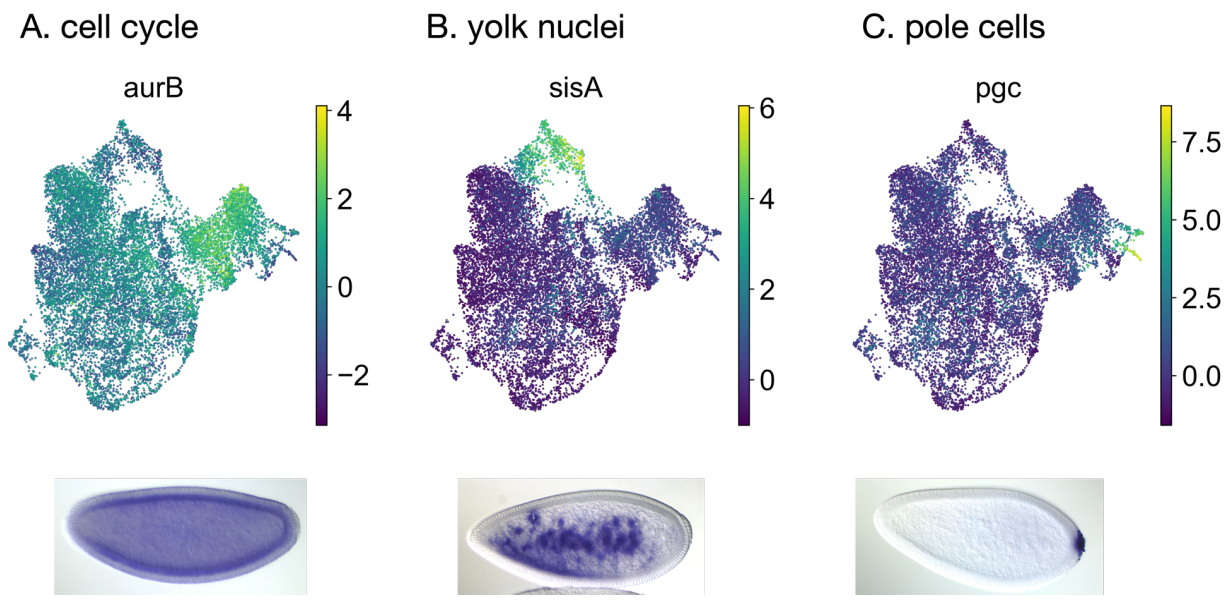

Figure 3.5 Clustering by cell cycle, yolk, and pole cell markers Figure depicts 2D umap representation of all nuclei colored by log(scVI normalized gene expression) for (A) aurB, a cell cycle marker, (B) sisA, a yolk nuclei marker, and (C) pgc, a pole cell marker. Representative *in situ* hybridizations for each gene (A, B, and C) are located underneath each respective UMAP.

Several studies show that transcription changes across the cell cycle, as shown by single-cell RNA-sequencing data, and present methods to account for the resulting variability in expression[106–108]. It is possible to batch correct the data accounting for cell cycle effects on gene expression; however, given nuclear division in the early *Drosophila* embryo occurs on the scale of minutes[109], I decided not to correct for this at first and examine expression of cell cycle genes. The nuclei do appear to somewhat cluster on *aurB* expression (see Figure 3.5.A). *aurB* is a ubiquitously expressed gene that is critical for chromatin condensation, among other functions in mitosis[110]. The nuclei that cluster with high *aurB* expression also have the highest percent mitochondrial expression (see Figure 3.4.C), which provides further assurance that these nuclei should be removed from the analysis.

Although I would not like to include yolk nuclei and pole cells in my analysis, nuclei that express a yolk nuclei marker cluster closely together (Figure 3.5.B). Similarly, nuclei that express a pole cell marker cluster closely together (Figure 3.6.B). I removed these clusters prior to further analysis; however, the fact that nuclei and the respective clusters retain a spatial identity provide me with confidence that I have filtered these data to the best of my ability and I can move forward with the analysis.

2D representation of early *Drosophila* embryonic nuclei

To generate the final 2D representation of the nuclei, I removed nuclei as indicated in the previous section from the nucleus x gene matrix and re-ran the clustering algorithms.

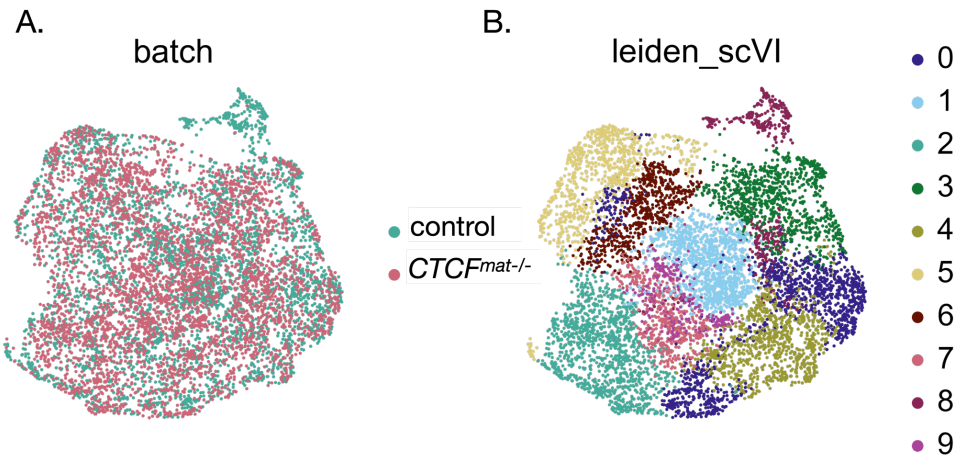

Figure 3.6 Two dimensional representation of early *Drosophila melanogaster* embryonic nuclei Figure depicts 2D UMAP representation of nuclei colored by (A) control (teal) and *dCTCF*$^{mat-/-}$ (pink) and (B) 10 nuclear clusters determined by the Leiden algorithm used within scVI.

I would not expect early embryonic nuclei to form strict clusters isolated from one another as I would with distinct cell types because at this point in development, cellularization and differentiation are just beginning. After batch correction and filtering, the two different samples widely overlap with one another (Figure 3.6.A) and although distinct clusters are not apparent in a reduced dimensional space, any differences between communities detected (Figure 3.6.B) by the clustering algorithms might suggest different nuclear identities.

At this point, I decided to do one last quality check and look at the same quality control metrics and gene expression as in Figures 3.4 and 3.5. If the filtering done above improved the quality of the data, I would expect the nuclei to no longer cluster on any of the features as mentioned previously.

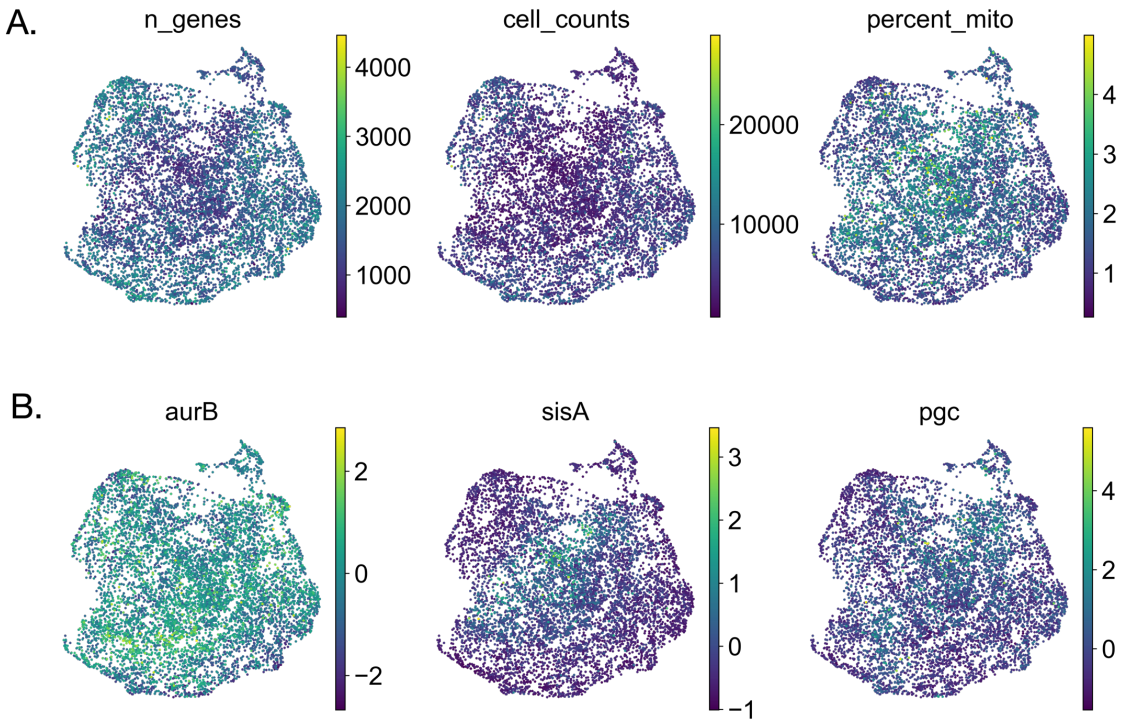

Figure 3.7 Nuclei no longer appear to cluster on undesirable qualities Figure depicts 2D UMAP representation of nuclei colored by (A) level of the same metrics as in Figure 3.4: number of genes expressed per nucleus, total UMI counts per nucleus, and percent mitochondrial expression within each nucleus, and (B) log(scvi normalized gene expression) in the same genes represented in Figure 3.5: *aurB*, *sisA*, and *pgc*.

As expected with adequate filtering, the nuclei no longer cluster on number of genes expressed per nucleus, UMI counts per nucleus, or percent mitochondrial expression (Figure 3.7.A). Additionally, the nuclei no longer appear to cluster on *aurB, sisA,* or *pgc* expression (Figure 3.7.B).

Marker gene analysis indicates that clusters have spatial identities

In order to ascertain whether or not each cluster represents a distinct spatial identity, I first examined the marker genes for each cluster. Although patterned gene expression is well characterized in the early *Drosophila* embryo, I did not necessarily expect gene expression in different clusters to be that distinct because I collected embryos prior to cellularization and differentiation.

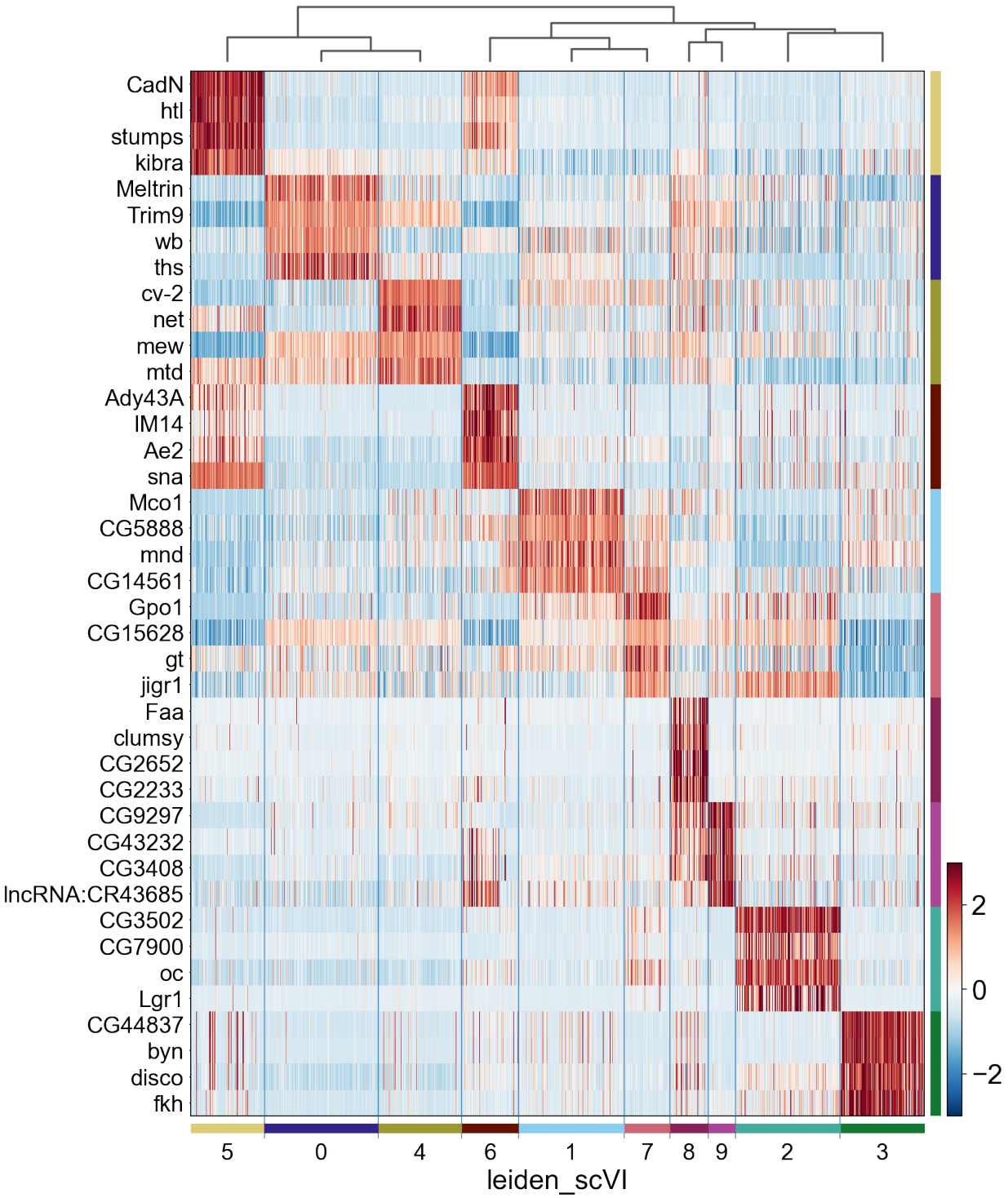

Figure 3.8 Marker gene heatmap Heatmap displays scaled gene expression of top four marker genes determined using a Wilcoxon signed-rank test of genes expressed in each cluster versus the rest of the clusters. Higher expression is shown in red, mean expression shown in white, and lower expression shown in blue.

Without even considering spatial patterning (or lack thereof), the marker genes representing each cluster are expressed in distinct patterns (Figure 3.8). Because these distinctions are so clear, I posited that these clusters retain spatial information from the position in the embryo where the

nuclei were located prior to dissociation. In order to do so, I looked through a list of marker genes that represent each cluster and examined gene expression patterns of representative *in situ* hybridizations in a public database.

0	1	2	3	4	5	6	7	8	9
Meltrin	*Mco1*	*CG3502*	*CG44837*	*cv-2*	*CadN*	*Ady43A*	*Gpo1*	*Faa*	*CG9297*
Trim9	*CG5888*	*CG7900*	*byn*	*net*	*htl*	*IM14*	*CG15628*	*clumsy*	*CG43232*
wb	*mnd*	*oc*	*disco*	*mew*	*stumps*	*Ae2*	*gt*	*CG2652*	*CG3408*
ths	*CG14561*	*Lgr1*	*fkh*	*mtd*	*kibra*	*sna*	*jigr1*	*CG2233*	*lncRNA: CR43685*
pyr	*CG34214*	*frma*	*rib*	*Samuel*	*CG9005*	*CG42808*	*hb*	*lncRNA: CR43700*	*CG12024*
hth	*geko*	*otk*	*lncRNA: CR43126*	*CheA84a*	*CG33725*	*CG32052*	*CREG*	*Sclp*	*CG13566*
Art7	*D*	*CG14204*	*tll*	*CG31523*	*if*	*hll*	*btd*	*CG13949*	*CG2225*
CG46442	*upd1*	*Graf*	*Fili*	*ush*	*lncRNA: CR45361*	*edl*	*Notum*	*CG15673*	*GstD8*
opa	*asRNA: CR4405*	*Gyg*	*bbg*	*CG14598*	*lbk*	*Cpr62Bb*	*lncRNA: CR44700*	*CG15617*	*Hsp67Bc*
raskol	*Cpr31A*	*Bili*	*mthl3*	*Cyp313b1*	*CG11357*	*CG8353*	*Socs44A*	*Pld*	*CG7530*

Table 3.1 Top 10 marker genes per cluster This table expands upon Figure 3.8, showing the top 10 marker genes for each cluster.

The Berkeley Drosophila Genome project has a collection of *in situ* hybridizations for over 8,000 genes in the *Drosophila melanogaster* embryo[98–100]. In order to determine if the distinct gene expression observed in Figure 3.8 corresponds to spatial patterning of the embryo, I searched this database for each gene present in Table 3.1, with representative *in situ* hybridizations shown in Figure 3.9.

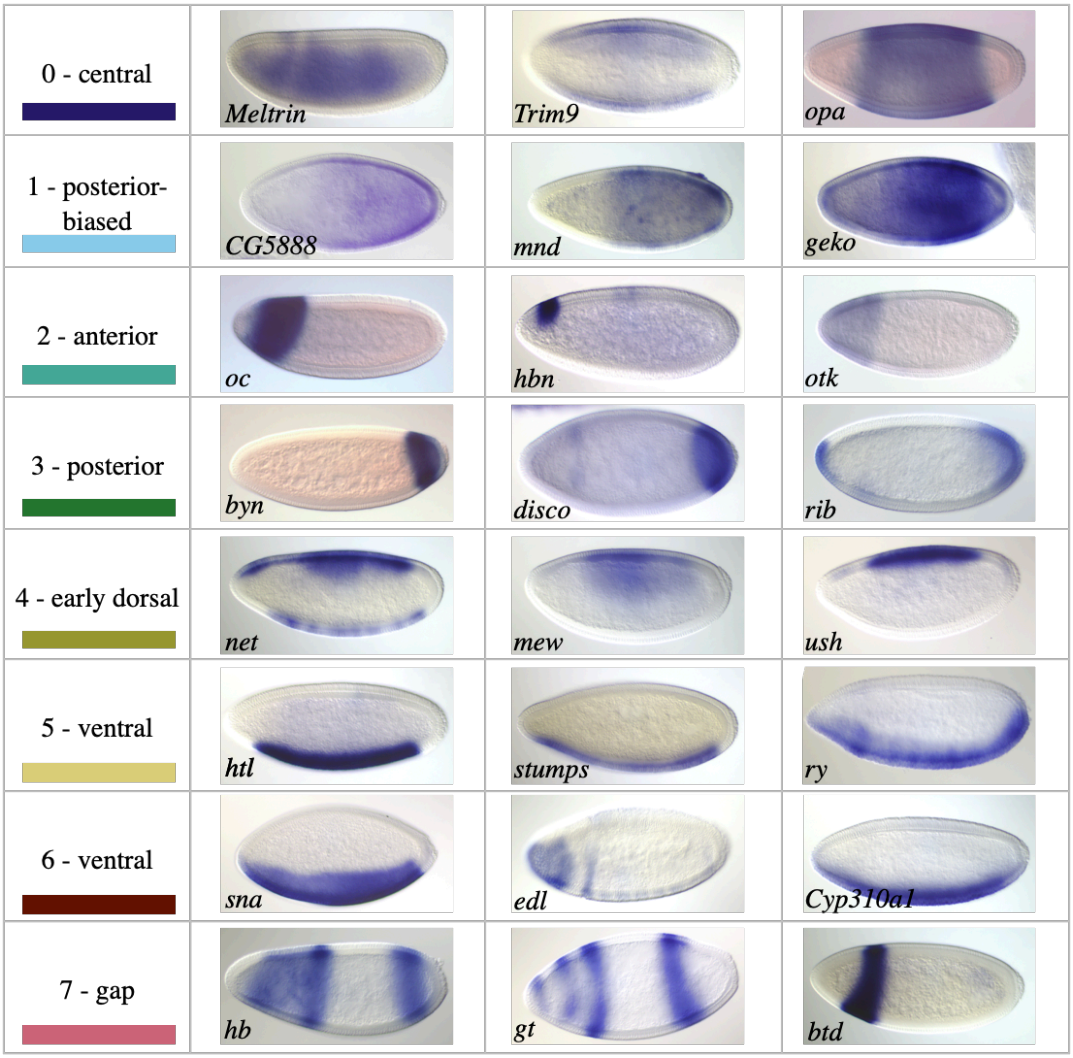

Figure 3.9 Representative *in situ* hybridizations of top marker genes for clusters 0-7 *In situ* hybridizations in row of the figure correspond to the cluster numbered on the left. Each image is of a stage 4-6 embryo[84]. If possible, I chose images from early to mid-stage 5 in order to show patterning close to the time point of my collection. On the right hand side, I have labeled each cluster according to the expression pattern of the representative images as well as additional images that I have not shown. The three genes listed underneath the identity of each cluster on the right hand side represent the genes probed for in the *in situ* hybridizations from left to right across the figure. I chose to omit clusters 8 and 9 because I could not confidently label either cluster. All *in situ* hybridizations are from the Berkeley Drosophila Genome Project[98–100].

Prior to looking at expression patterns of the marker genes in Figure 3.8 and Table 3.1, I was not sure if the clusters correspond to distinct regions of the embryo. Fortunately, I was able to confidently determine the region the nuclei were in prior to dissociation for clusters 0-7 (see Figure 3.9) where three or more genes in the top 10 corresponded to the regions listed on the right. All axes of the embryo are represented by the regions listed in Figure 3.9; however, I noticed that the marker genes for cluster 4 tended to express on the dorsal side early with ventral expression appearing later. Because I cannot strictly label cluster 4 as dorsal, I decided to call it early dorsal. Additionally, I was unable to find a consistent pattern of expression for the marker

genes in clusters 8 and 9, even looking beyond the top 10 marker genes, thus I cannot confidently say if these clusters represent nuclei from certain regions of the embryo.

With the majority of clusters representing spatial regions of the embryo, I then wanted to know if genes are differentially expressed upon loss of maternal dCTCF in order to determine if gene expression is differentially affected across different regions of the embryo. Whether or not this is the case for dCTCF, the establishment of this method and the ability to assign nuclei to spatial regions of the embryo opens many doors to investigate other factors important in the regulation of gene expression in the future.

Differential expression analysis in individual clusters yields differentially expressed genes not captured in bulk

Differential expression analysis is commonly done in RNA-sequencing experiments in order to find genes that are up or down-regulated following a perturbation. In order to determine if genes are differentially expressed spatially upon loss of maternal dCTCF, I conducted differential expression analysis in bulk and in individual clusters.

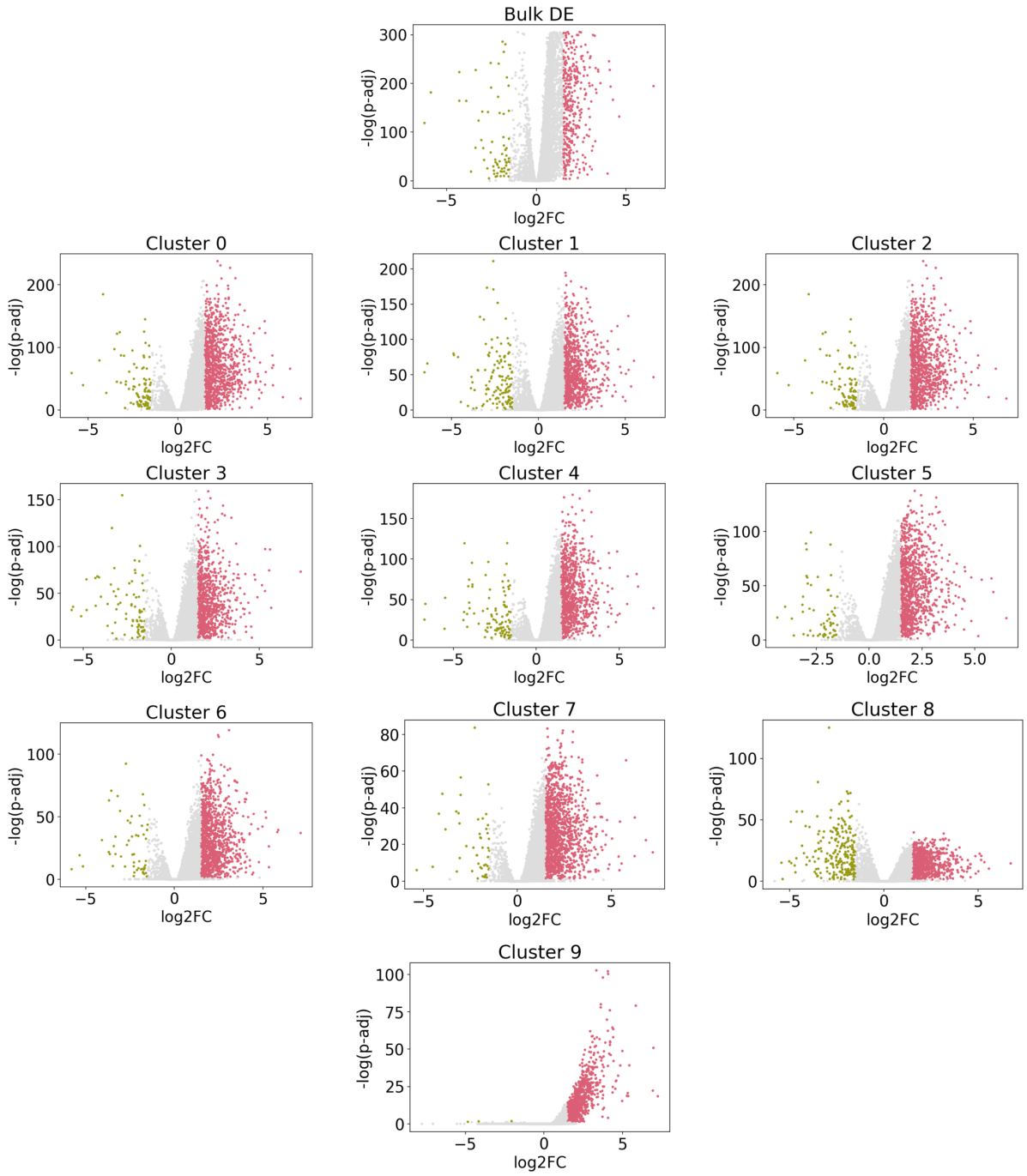

Figure 3.10 Differential expression in bulk and individual clusters Figure depicts volcano plots of log2FC (log2(fold-change)) by the log of the adjusted p-value (p-adj) for differential expression calculated in bulk (top middle) and in individual clusters. I considered a gene significantly expressed for an absolute value of log2FC >= 1.5 and p-adj < 0.05. Significantly down-regulated genes upon loss of maternal dCTCF are green, significantly up-regulated genes are in pink. Non-significantly differentially expressed genes are in light gray.

As shown in Figure 3.10, the majority of differentially expressed genes in bulk and individual clusters are up-regulated upon loss of maternal dCTCF. This observation might have implications on dCTCF function in enhancer-blocking or bridging; however, this remains unclear without repeating this experiment and conducting follow-up studies that confirm (or refute)

altered expression in specific genes. I also noticed that fewer genes are differentially expressed in bulk relative to the clusters, which then led me to ask how many differentially expressed genes the bulk analysis has in common with the individual clusters.

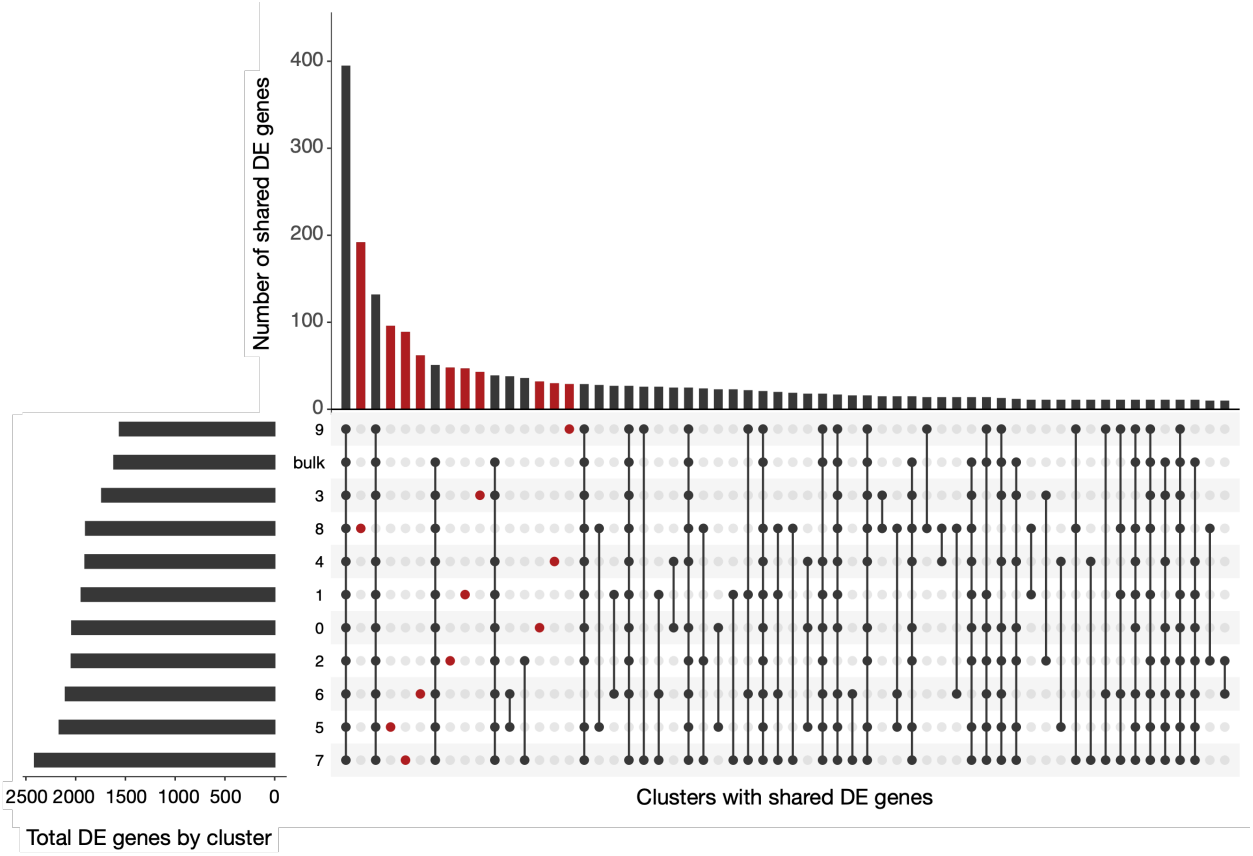

Figure 3.11 Intersecting differentially expressed genes between bulk and individual clusters
Figure depicts an UpSet plot for visualizing shared traits between many conditions. The horizontal bar plot (left) is sorted by the total number of differentially expressed (DE) genes within individual clusters or the bulk analysis. Genes are considered differentially expressed if the corrected p-value (p-adjusted * number of conditions, 11). The vertical bar plot (top) represents the number of shared DE genes for the conditions indicated below. Connected dots (black) represent groups of genes that are differentially expressed in the indicated groups. Genes that are only DE in one group are colored in red.

As shown in Figure 3.11, differentially expressed genes shared between all clusters and bulk analysis represent the largest category of overlap. With that being said however, a significant number of differentially expressed genes are only found in single clusters (shown in red). Additionally, some differentially expressed genes are shared between multiple clusters but not in bulk (black connected dots without bulk filled in). Because most of these clusters represent spatial regions of the embryo, these results suggest that loss of dCTCF may differentially affect gene expression in physical space.

Differential expression of key patterning genes not detected in bulk

In order to determine whether or not spatially patterned genes are present in the differentially expressed genes that are not intersecting, or those that intersect (not with bulk) in Figure 3.11, I examined the overall expression of each gene within each nucleus separated by control and *dCTCF*$^{mat-/-}$ conditions in each cluster and in bulk. I found that many patterned genes are differentially expressed in clusters, but not in bulk; however, I decided to limit the number of genes plotted to two per spatial region for the sake of brevity.

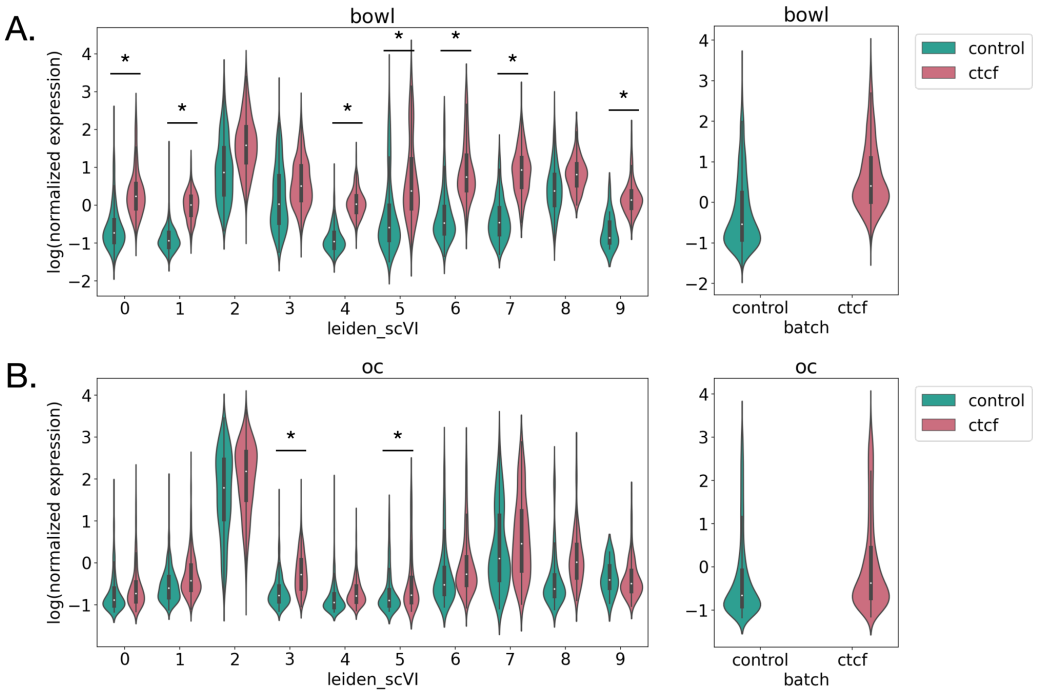

Figure 3.12 Expression of two anterior patterned genes in each cluster and in bulk Figure depicts the log of scVI normalized expression in each cluster (left) and in bulk (right) for two anterior patterned genes, (A) *bowl* and (B) *oc,* in control (teal) and *dCTCF*$^{mat-/-}$ (pink). Asterisk (*) indicates significant differential expression where the adjusted p-value < .05 and the absolute value of log2FC >= 1.5.

Interestingly, *bowl* is differentially expressed in several clusters, but not in the anterior cluster where its primarily found (cluster 2, see Figure 3.12.A). *bowl* expression in cluster 2 and in bulk does appear up-regulated upon loss of maternal dCTCF; however, these are not considered significant differences because the log2FC is less than 1.5. Given that *bowl* is significantly up-regulated in several clusters, even though not the anterior cluster, these results indicate that loss of dCTCF results in increased *bowl* expression.

Another anterior gene, *oc,* is differentially expressed in two clusters, again not in the anterior cluster (2, See Figure 3.12.B). However, in this case *oc* is up-regulated in cluster 3 which represents posterior nuclei, and cluster 5 which represents ventral nuclei. The marker genes representative of cluster 2 includes *oc* (see Table 3.1 and Figure 3.9). Other *in situ* hybridizations by the Berkeley Drosophila Genome Project do not show expression of *oc* in the posterior region of the embryo[98–100]. This might indicate misexpression of *oc* in the posterior; however, without further experiments, such as single-molecule RNA FISH of nuclear *oc* RNA, I cannot confirm that this is a biological finding and not an artifact.

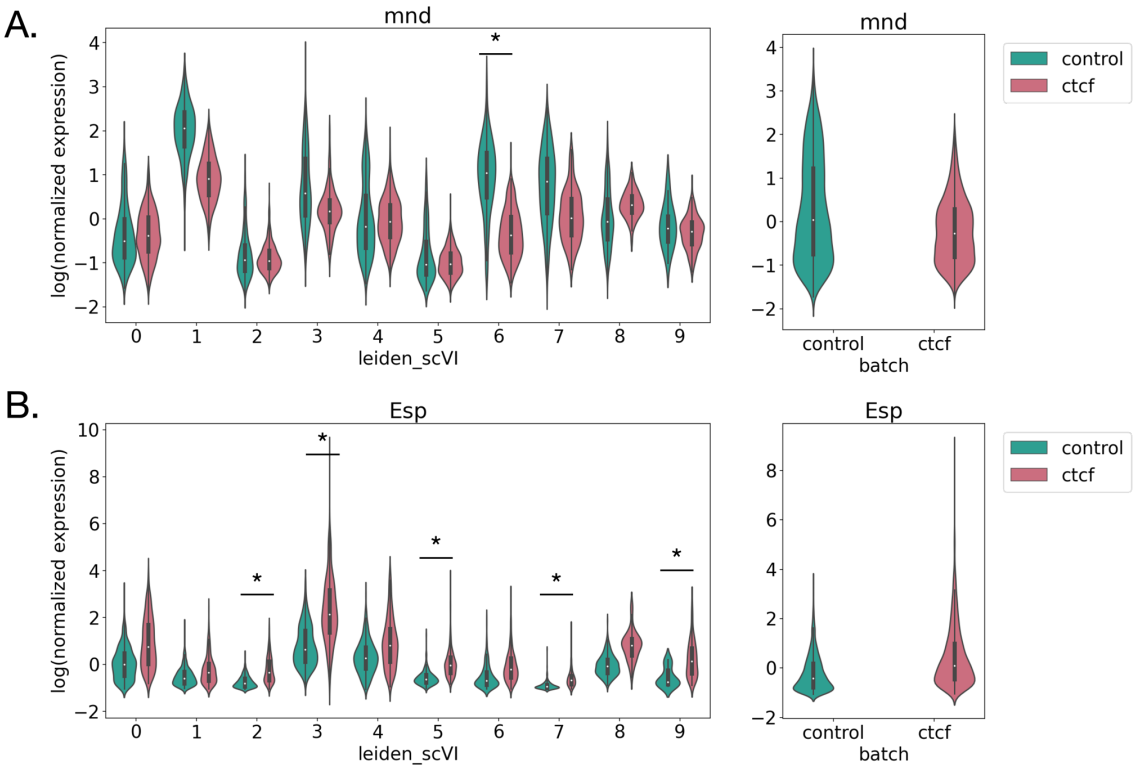

Figure 3.13 Expression of two posterior patterned genes in each cluster and in bulk Figure depicts the log of scVI normalized expression in each cluster (left) and in bulk (right) for two posterior patterned genes, (A) *mnd* and (B) *Esp*, in control (teal) and *dCTCF*$^{mat-/-}$ (pink). Asterisk (*) indicates significant differential expression where the adjusted p-value < .05 and the absolute value of log2FC >= 1.5.

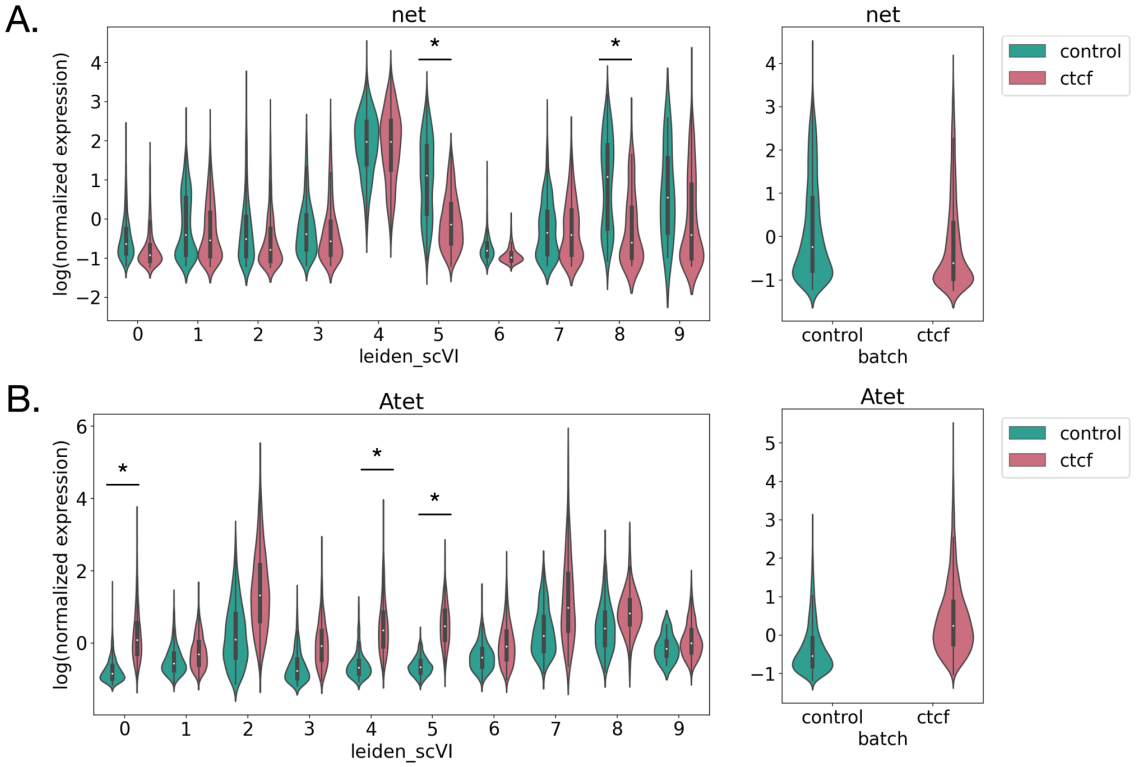

Figure 3.14 Expression of two dorsal patterned genes in each cluster and in bulk Figure depicts the log of scVI normalized expression in each cluster (left) and in bulk (right) for two dorsal patterned genes, (A) *net* and (B) *Atet*, in control (teal) and *dCTCF*$^{mat-/-}$ (pink). Asterisk (*) indicates significant differential expression where the adjusted p-value < .05 and the absolute value of log2FC >= 1.5.

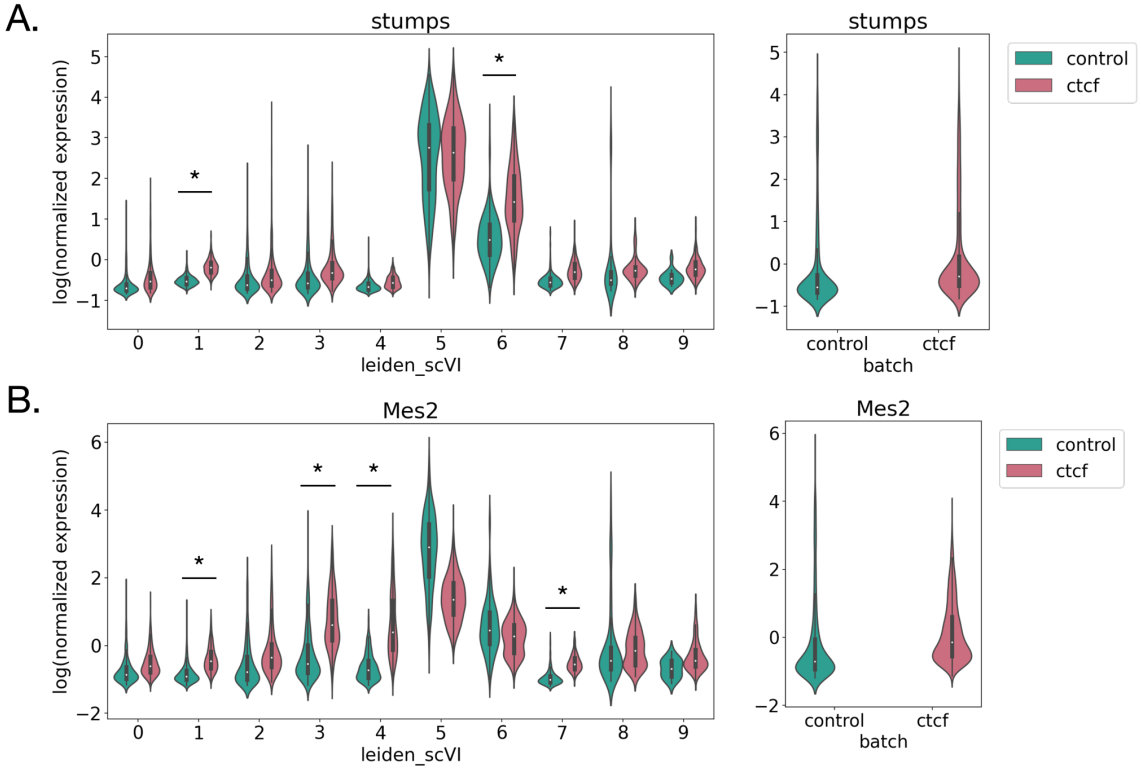

Figure 3.15 Expression of two ventral patterned genes in each cluster and in bulk Figure depicts the log of scVI normalized expression in each cluster (left) and in bulk (right) for two ventral patterned genes, (A) *stumps* and (B) *Mes2*, in control (teal) and *dCTCF*$^{mat-/-}$ (pink). Asterisk (*) indicates significant differential expression where the adjusted p-value < .05 and the absolute value of log2FC >= 1.5.

Similar observations can be made for posterior (Figure 3.13), dorsal (Figure 3.14), and ventral genes (Figure 3.15) where patterned genes are differentially expressed in clusters, but not in bulk. I assume that changes in lowly expressed genes are more likely to be significant, as the overall level would not have to change as much in lowly expressed genes as in highly expressed genes. As such, I would consider observations regarding the number of differentially expressed genes cautiously and I decided to examine average gene expression across differentially expressed genes and non-differentially expressed genes.

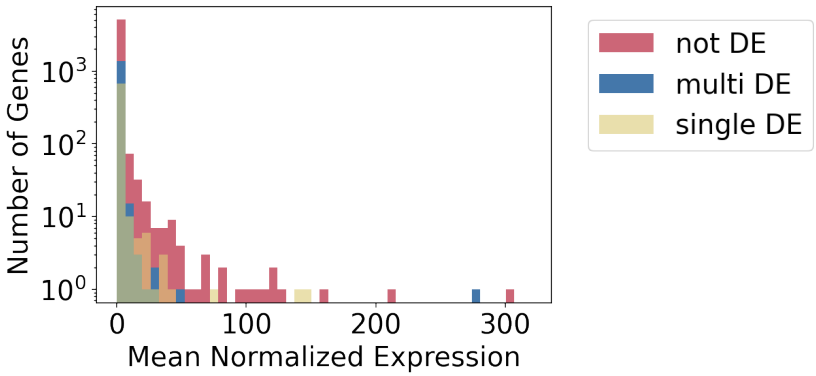

Figure 3.16 Mean normalized expression of differentially expressed genes follows the same distribution of the mean of non-differentially expressed genes Figure shows a histogram of the average gene expression of genes differentially expressed as defined by the results from Figure 3.11 in one group (yellow), differentially expressed in multiple groups (blue), and not differentially expressed (red). Each count on the y-axis represents a single gene.

Although the relative gene expression for the differentially expressed patterned genes above (Figure 3.12-3.15) is low, the distribution of average expression in differentially expressed genes follows a similar distribution of expression in non-differentially expressed genes (see Figure 3.16). The fact that gene expression is skewed towards the lower side is not surprising, considering the embryos were collected during zygotic genome activation as gene expression is established. Altogether, the results shown above demonstrate that single-nucleus RNA-sequencing allows for the detection of gene expression changes in spatial regions across pre-cellularized embryos.

Discussion

Here, I demonstrate the use of single-nucleus RNA-sequencing to detect changes in spatial gene expression upon loss of dCTCF. Single-nucleus RNA-sequencing is incredibly noisy, based on the fact that I began with over 200,000 nuclei in each condition and ended up with a little over 8,000 nuclei total in the final analysis. After initial filtering, I was able to demonstrate that this technique is highly specific for known types of nuclei (yolk and pole cell, Figure 3.4). After removing the clusters corresponding to yolk nuclei and pole nuclei, I was able to show that nuclear gene expression clusters on spatially-patterned genes depending on the origin of the nucleus by examining the wild-type expression pattern of representative marker genes from each cluster (Figure 3.8, Figure 3.9, and Table 3.1). Then, I wanted to know whether differentially-expressed genes are differentially expressed between individual clusters and bulk data.

I found that many genes are differentially expressed across all clusters and in bulk; interestingly however, many genes are only differentially expressed in one cluster and not in bulk (Figure 3.11). In several cases, spatially-patterned genes that are not differentially expressed in bulk are in fact differentially expressed in one or more clusters (Figures 3.12-3.15). This finding highlights the utility of single-nucleus RNA-sequencing over RNA-sequencing in bulk and generates a list of potential candidates for follow-up studies. I did notice that the majority of differentially expressed genes have low means which may be a concern; however, the means of

differentially expressed genes are not drastically different from non-differentially expressed genes (Figure 3.16).

Whether or not the changes in expression that I observed have implications in embryonic development is not clear without further investigation, such single-molecule FISH of specific nuclear RNAs to quantify the number of molecules present in each nucleus. With that being said, I have shown in general that single-nucleus RNA-sequencing can be used to examine patterned gene expression and differential expression across spatial regions of the embryo and this can be used to generate candidates for further analysis. The ability to do this exponentiates the number of genes and nuclei that we can examine under various perturbations during early *Drosophila* embryonic development.

Chapter 4: Towards a single-cell RNA-sequencing split-pool barcoding strategy using catalytic hairpins

Abstract

Despite an exponential increase in the number of cells included in single-cell RNA-sequencing experiments since its inception, current technologies either rely on expensive equipment, large quantities of proprietary reagents, or require procedures that result in sample loss. In this chapter, I describe a series of experiments conducted in an effort to improve upon existing methods of barcoding cells in single-cell RNA-sequencing experiments. With my collaborators, I have designed and tested multiple versions of primer exchange reaction (PER) catalytic hairpins in an effort to solve the existing barcoding problem. The ideal barcoding strategy would not rely on expensive reagents and limit sample loss during washing steps, and PER solves this problem using simple oligos that form hairpins designed to specifically extend a complementary sequence. By designing these hairpins containing a barcode as well as a universal sequence that is specific to the round prior, PER would completely eliminate sample washing in between rounds of barcoding. I attempted two strategies to capture each end of the RNA, termed 5' and 3' PER. PER barcoding is highly efficient at extending single-stranded DNA *in vitro;* however, barcoding is much less efficient following reverse transcription whether intended to capture 5' or 3' RNA sequences. Fully barcoded sequence is visible upon overexposure of gel images, but fully barcoded cDNA remains undetectable in pilot sequencing experiments. Even so, with further optimization PER remains a viable strategy for barcoding cells *in situ* from the high efficiency observed under certain conditions *in vitro*.

Introduction

Single-cell and single-nucleus RNA-sequencing are becoming promising methods for our understanding of gene expression during embryonic development across many species[71,91,111–113] (see Chapter 3). However, much of this work was conducted using either proprietary or homemade microfluidic devices. Many labs may not have access to this equipment or technology, nor have the expertise to build a rig for themselves. With that being said, we need to develop methods that do not require microfluidics, or other specialized equipment, in order to drive the single-cell field forward.

Combinatorial indexing (from sci-RNA-seq), or split-pool barcoding (from SPLiT-seq) are powerful single-cell RNA-sequencing barcoding methods that yield on the order of tens of thousands to over 100,000 cells by conducting barcoding across multiple rounds using PCR plates as opposed to microfluidics[69,91]. As I discussed in Chapter 1 of this dissertation, in terms of the number of cells sequenced, this is exponentially greater than the manually-isolated cells in the first ever scRNA-seq experiment[66]. However, both sci-RNA-seq and SPLiT-seq are not without their respective limitations.

SPLiT-seq relies on ligation to add barcodes, thus the samples must be washed in between each round of ligation which results in sample loss. Anecdotally amongst myself and other researchers

I know who have conducted any single-cell experiment, isolating cells or nuclei from biological samples is the greatest challenge, especially when no protocols exist for whatever organism is used. This is also true for perturbation experiments, i.e. genetic mutants, where samples may not be healthy and produce a limited amount of starting material. Sci-RNA-seq bypasses the ligation problem using barcoded reverse transcription (RT) primers, Illumina TDE1, and PCR primers. However, because TDE1 (commonly known as Tn5) is a proprietary reagent and the 96-well barcoded TDE1 is not publicly available, this method is not feasible for most labs to reproduce. Making barcoded Tn5 in-house is possible[114], but not every lab is equipped to do so. With these limitations in mind, I believe improvements to single-cell barcoding are necessary to drive the cost of single-cell experiments down while simultaneously increasing the availability of single-cell methods to more scientists.

In an effort to improve single-cell barcoding, I began a collaboration with Jase Gehring, Taleen Dilanyan, and Lior Pachter at CalTech. Single-cell barcoding is more simply described as extension of DNA. Kishi et al (2018) published work on primer exchange reaction (PER) cascades where a catalytic hairpin that recognizes a single-stranded DNA template is able to extend a single-stranded template with a new sequence to be specifically recognized by another catalytic hairpin[115]. The authors presented the use of these hairpins for DNA origami, RNA degradation, among other methods; however, we posited that we could adapt their design for single-cell barcoding by adding a 5-bp barcode in between the landing sites present in each hairpin (see Figure 4.1). Together, we designed two single-cell RNA-sequencing approaches utilizing these catalytic hairpins inspired by both sci-RNA-seq and SPLiT-seq (see Figures 4.2-4.4).

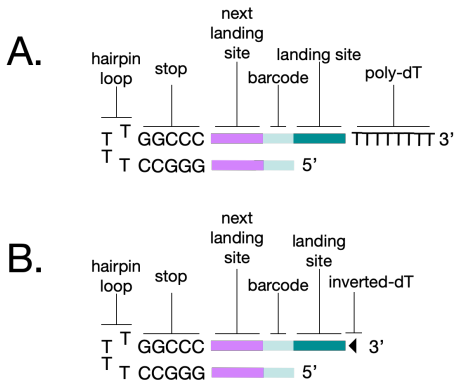

Figure 4.1 Catalytic hairpins designed for use in split-pool barcoding for single-cell RNA-sequencing Figure depicts two versions of a catalytic hairpin used to extend and barcode single-stranded DNA. Both hairpins contain the same features with the exception of the 3' tail, (A) ends in a stretch of 8 Ts and (B) ends in an inverted-dT. Describing the features from left to right as presented above, each hairpin contains a sequence of 4 Ts that do not base pair such that a hairpin is formed. Kishi et al presented multiple options for stop sequences, but I chose to use the absence of dGTPs in solution. Once the polymerase reaches the Cs in the stop sequence, the hairpin will fall off and continue to the next substrate. The 'next landing site' is a sequence that is added to the end of the single-stranded DNA substrate that is recognized by the next hairpin. As these are designed for barcoding, each individual hairpin adds 5 bases unique to the hairpin itself. The 'landing' site is a sequence complementary to the end of the single-stranded DNA substrate such that the hairpin can bind and a polymerase will extend the substrate

using the rest of the hairpin as a template. Finally, both the (A) poly-dT region and (B) inverted-dT at the 3' end of each hairpin serve to prevent the hairpin itself from extension by a polymerase.

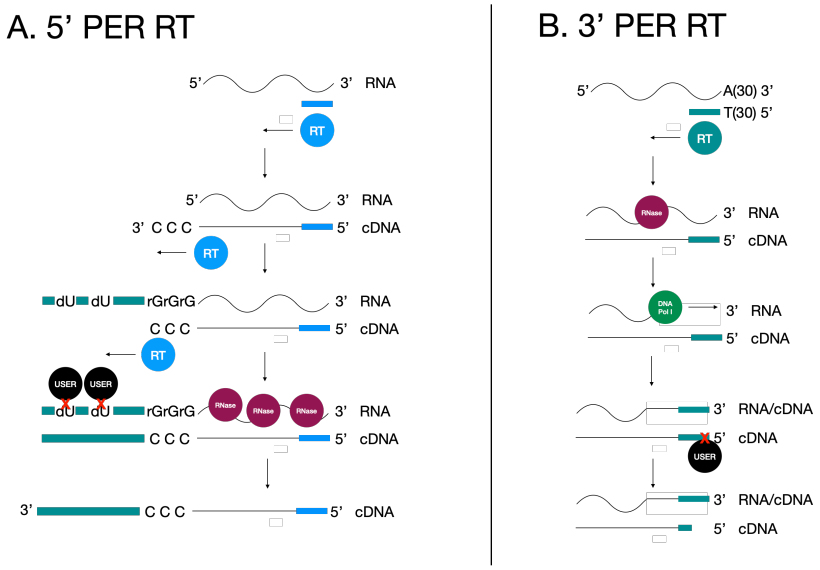

Figure 4.2 Diagram of RT preceding PER barcoding Figure depicts reactions that occur during RT for (A) 5' PER and (B) 3' PER. These are named 5' or 3' PER for the end of the RNA that will eventually be represented in eventual sequencing data. For (A) 5' PER, the reverse transcriptase (blue) is primed with an oligo that contains a PCR handle for downstream amplification. Once the reverse transcriptase (blue) reaches the 5' end of the RNA, the terminal-transferase of the reverse transcriptase (blue) adds three Cs on the 3' end of the cDNA. This is used as a template for the template-switching oligo (TS) to extension of the cDNA in the next step. USER enzyme (black) then cleaves the dUs in the TSO while RNaseH (maroon) cleaves RNA/DNA hybrids. The resulting single-stranded cDNA contains a sequence complementary to the TSO, which serves as a landing site for the first PER barcoding reaction, and a PCR handle on the 5' end for downstream amplification. For (B) 3' PER, a different reverse transcriptase (teal) is primed with an oligo-dT(30) containing a PCR handle for downstream amplification. After RT, a mild RNaseH (maroon) treatment generates few nicks in the RNA/DNA hybrid. These nicks serve as a primer for second-strand synthesis via DNA Pol I (green). USER enzyme (black) cleaves dUs in the RT primer to generate a single-stranded overhang on the 3' end of the newly generated cDNA, which will serve as the first landing site for PER.

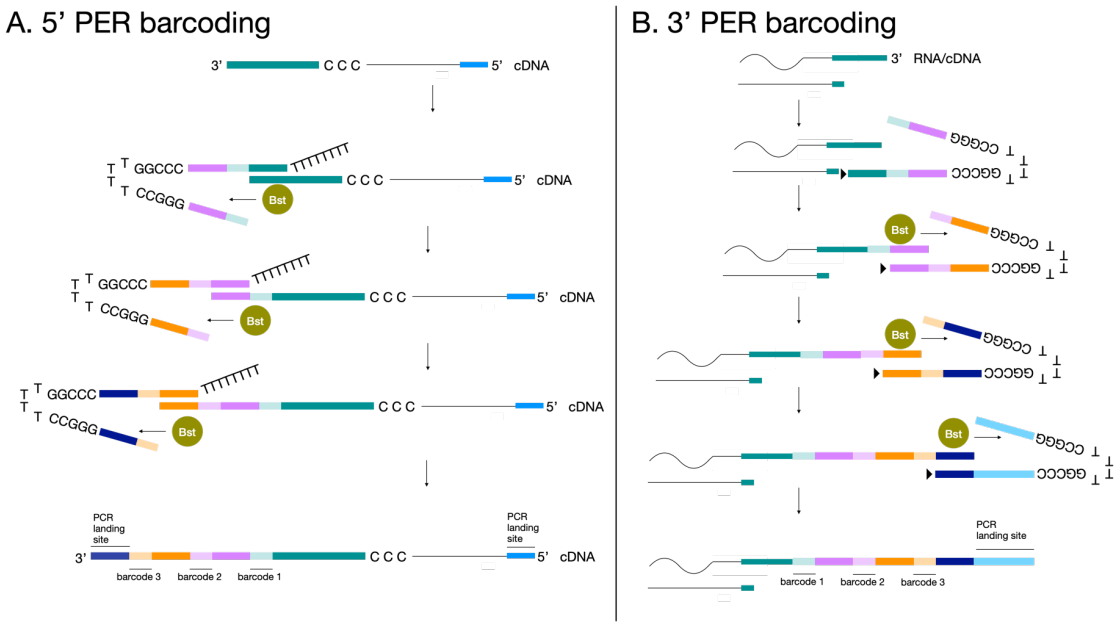

Figure 4.3 Diagram of PER barcoding Figure outlines PER barcoding with (A) 5’ PER and (B) 3’ PER. On the left I have shown the use of poly-dT hairpins (Figure 4.1.A) and on the right the inverted-dT hairpins (Figure 4.1.B); however, the two may be used interchangeably. For (A) 5’ PER, the first hairpin binds to the extended sequence complementary to the TSO described in Figure 4.2, and Bst polymerase (green) displaces the hairpin double-stranded region and extends the cDNA with a 5 base barcode (light teal) and a new landing site (pink) for the next hairpin. This hairpin extension is repeated two more times with different barcoded hairpins, where each round of barcoding is specific to the one before (indicated by the different colors added in each round). The universal sequence added by the last hairpin (dark blue) serves as a landing site for a downstream PCR primer. For (B) 3’ PER barcoding, the first hairpin binds to the overhang while Bst polymerase (green) displaces the hairpin double-stranded region and extend the overhang by adding a 5 bp barcode (light teal) and landing site for the next hairpin. This, similarly to 5’ PER, occurs two more times with the exception of an additional round of PER to add a longer PCR landing site, which will benefit the downstream PCR reaction.

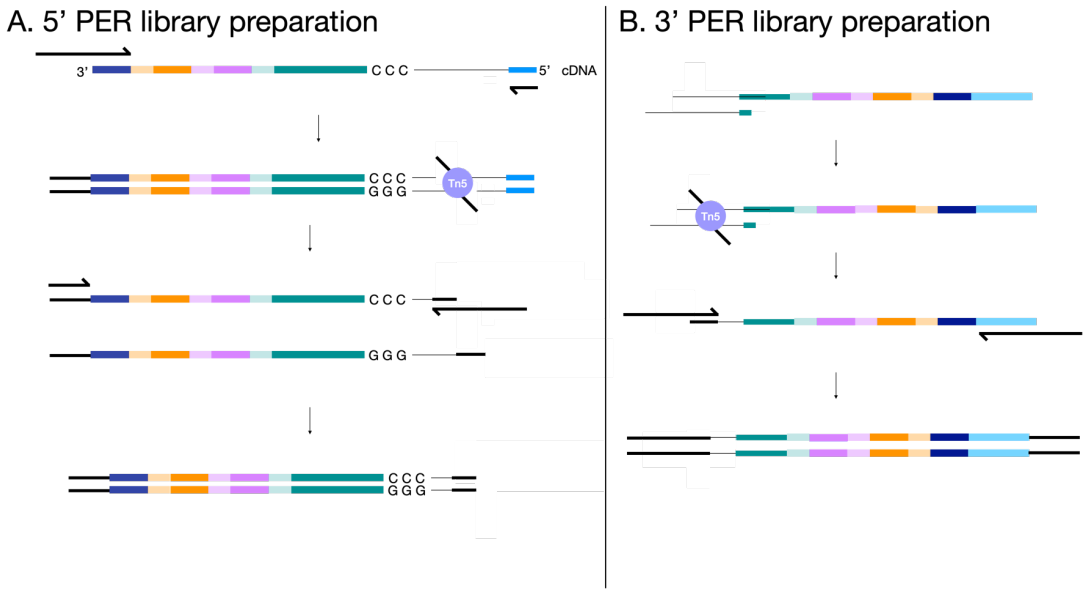

Figure 4.4 Diagram of library preparation following PER barcoding Figure represents library preparation for next-generation sequencing following PER barcoding. For library preparation, the initial product must be double-stranded, so the first step of the reaction differs between (A) 5' PER and (B) 3' PER. For (A) 5' PER, an initial round of five PCR cycles is plenty to generate double-stranded cDNA. After this point, the protocol for (A) 5' PER and (B) 3' PER are the same. Tn5 loaded with Illumina adapters inserts itself into the double-stranded cDNA and inserts an adapter at the end of each strand. After this tagmentation, the libraries are amplified to generate sufficient amounts of material for sequencing.

As shown above in Figures 4.2-4.4, our PER hairpin design will allow us to barcode and sequence either end of cDNA molecules following reverse transcription. Using barcoded PER hairpins with three or more rounds of split-pool barcoding in 96 well plates (see Chapter 1, Figure 1.2), we will be able to sequence RNA from tens of thousands, to hundreds of thousands of cells without having to wash the samples in between rounds of barcoding. Ultimately, development of this method has the potential to improve access to single-cell technologies as the hairpins are inexpensive DNA oligos and we utilize common laboratory reagents and enzymes to complete the process. In addition, the design of the barcoding hairpins eliminates the need for cell washing, which reduces sample loss.

In this chapter, I describe a set of experiments demonstrating the efficiency of the above described PER hairpins *in vitro*. I found that our hairpin designs efficiently extend a single-stranded DNA template *in vitro*, paralleling the efficiency demonstrated by Kishi et al (2018). However, this extension was much less efficient at extending cDNA following reverse transcription. Several strategies may mitigate some of this inefficiency, but not to the full extent of experiments without reverse transcription. The presence of ribosomal RNAs in the reaction remains a concern during sequencing, thus I tried multiple strategies to mitigate this issue as well. In the end, pilot sequencing experiments did not yield properly barcoded cDNA; however, this remains an ongoing project as additional steps of the library preparation process need further optimization. Barcoding with PER hairpins remains a promising strategy for use in single-cell RNA-sequencing as our design is highly efficient *in vitro*. With that being said, further work is necessary to establish this as a method and this project is ongoing.

Methods

Extension of single-stranded DNA in vitro

Unless otherwise indicated, I performed all experiments extending single-stranded DNA *in vitro* in 50 uL reactions with the following specifications: 5 uL 10x ThermoPol, 5 uL 10x Bst Polymerase, 5 uL 100 mM MgSO4, 5 uL 1 mM dHTP, 5 uL 100 mM spermidine, 0.5 uL 100 uM Cy5 oligo, 5 uL hairpin (400 nM of each hairpin) mix, 19.5 uL H2O. I incubated the samples in a thermocycler for 4 hours at 37°C and heat inactivated the reaction at 80°C for 20 minutes.

Oligo	Sequence
AAH1-1	ACCTCCAAACCTCAGGGCCTTTTGGCCCTGAGGTTTGGAGGTAGATCGGTATTTTTTTT
AAH2-1	ACTTTCAAACCTCAGGGCCTTTTGGCCCTGAGGTTTGAAAGTAGATCGGTATTTTTTTT
AAH3-1	AACATCAAACCTCAGGGCCTTTTGGCCCTGAGGTTTGATGTTAGATCGGTATTTTTTTT
AAH4-1	AACTCCAAACCTCAGGGCCTTTTGGCCCTGAGGTTTGGAGTTAGATCGGTATTTTTTTT
AA196	/5Cy5/CGCTCTACCGATCT

Table 4.1 Oligos used for extension of single-stranded DNA *in vitro* with poly-dT tail

Oligo	Sequence
AA268	ACCTCCAAACCTCAGGGCCTTTTGGCCCTGAGGTTTGGAGGTAGATCGGTA/3InvdT/
AA269	ACTTTCAAACCTCAGGGCCTTTTGGCCCTGAGGTTTGAAAGTAGATCGGTA/3InvdT/
AA270	AACATCAAACCTCAGGGCCTTTTGGCCCTGAGGTTTGATGTTAGATCGGTA/3InvdT/
AA271	AACTCCAAACCTCAGGGCCTTTTGGCCCTGAGGTTTGGAGTTAGATCGGTA/3InvdT/
AA196	/5Cy5/CGCTCTACCGATCT

Table 4.2 Oligos used for extension of single-stranded DNA *in vitro* with inverted-dT

Oligo	Sequence
AA247	CCACTCACCTCCTTTAAACTGGGCCTTTTGGCCCAGTTTAAAGGAGGTGAGTGGGAGTGTTAGTTTTTTTT
AA250	CCACTCCTTTAAACTGGGCCTTTTGGCCCAGTTTAAAGGAGTGGGAGTGTTAGTTTTTTTT
AA253	ACATCACTTTAAACTGGGCCTTTTGGCCCAGTTTAAAGTGATGTGAGTGTTAGTTTTTTTT
AA256	ACATCATTCCACTTTAAACTGGGCCTTTTGGCCCAGTTTAAAGTGGAATGATGTGAGTGTTAGTTTTTTTT
AA4-UMI	CTTTAAACTGGGCCTTTTGGCCCAGTTTAAAGNNNNNNNNGAGTGTTAGTTTTTTTT
AA4-U	CTTTAAACTGGGCCTTTTGGCCCAGTTTAAAGGAGTGTTAGTTTTTTTT

Table 4.3 Oligos used to extend bases added with the 4th hairpin

Reverse transcription with template switching in vitro

Unless otherwise indicated, I performed all RT experiments as follows on ice. To anneal the reverse transcription (RT) primer with the template, I added 0.25 uL of an RNA oligo, 0.1 uL of 100 uM Cy5 RT primer, 1 uL 10 mM dNTP, and 4.775 uL H2O to a PCR tube and mixed 10 times by pipetting. In a separate tube I combined 2.5 uL Template Switching RT Buffer (NEB),

0.375 uL of the indicated 100 uM template switching oligo, and 1 uL Template Switching RT Enzyme Mix (NEB) and mixed 10 times by pipetting. I added the 3.875 uL RT reaction mix to the 6.125 uL annealing mix above and mixed 10 times by pipetting. I briefly spun down the tubes, then incubated each sample for 90 minutes at 42°C and heat inactivated for 5 minutes at 85°C.

Oligo	Sequence
AA 207	rGrUrGrGrCrUrArGrArArUrUrGrUrArGrUrCrArUrArArUrUrArArGrGrCrG rCrArA
AA TSO2	AGAdUCGGdUAGAGCGTCGTGTArGrGrG
AA 221	CGGAGATGTGTATAAGAGACAGNNNNNN

Table 4.4 Oligos used for template-switching reverse transcription *in vitro*

TBE-PAGE

I poured my own 15% TBE-PAGE gels by adding 10g Urea and 2 mL 10X TBE to a 50 mL conical, then filled to 20 mL with H2O. In the hood, I added 160 uL 10% ammonium persulfate and 20 uL TEMED mixed quickly. I pipetted the mix into an assembled gel caster using a serological pipette, then waited for the mixture to solidify prior to running the gel.

To 5 uL of each sample, I added 5 uL of 2x formamide loading buffer. I heated each sample to 90°C for 5 minutes and snap cooled on ice. While the samples were heated, I microwaved 800 mL X TBE for 3 minutes and poured this into the gel box. I ran the gels at 200 V for 30 minutes, then stained in 3X GelGreen for 10 minutes prior to imaging Cy5 and GelGreen channels on a ChemiDoc Imaging System.

Bulk embryo RNA extraction

Rather than pilot these experiments in nuclei, I decided to simplify pilot experiments by extracting bulk RNA from overnight OreR (wild-type) embryo collections. I placed 6-10 vials of OreR flies in a medium collection cage and fed the cages with yeast paste made from Red Star yeast pellets and water, spread on molasses plates. Once the flies acclimated to the cage after 3-4 days, I fed the cage at the end of the day and collected several hundred embryos overnight. I proceeded with the RNA extraction described for single-embryo RNA extractions in Chapter 2. In this case however, I did not sort embryos and I extracted the RNA in bulk.

5' PER

I performed 5' PER by starting with *in vitro* reverse transcription as described above, with the exception of using bulk RNA extracted from embryos instead of an RNA oligo. I also designed a non-Cy5 RT primer compatible with PCR required in the library preparation step. Prior to barcoding, I incubated each sample with 2 uL rSAP and 1 uL RNaseH per 10 uL reaction. If I used a single-stranded TSO containing a dU, I also added 1 uL USER enzyme per 10 uL reaction.

Following template switching, I conducted the barcoding step according to the extension of single-stranded DNA in vitro method described above with the exception of using 10 uL of the RT reaction as the template and adjusting the amount of H2O accordingly. After barcoding, I proceeded with library preparation as described below.

3' PER

For 3' PER libraries, I performed RT as previously described, with modifications[91]. To anneal RT primer to RNA, I combined 1 uL of RNA template, 0.5 uL of 100 uM RT primer, 0.5 uL of 10 mM dNTP, and 4.5 uL H2O. After mixing 10 times by pipetting, I centrifuged the samples briefly and incubated the samples for 5 minutes at 55°C and snap cooled on ice. I prepared the RT mix by combining 2 uL of 5x Superscript IV Buffer (ThermoFisher), 0.5 uL of 100 mM DTT, 0.5 uL of Superscript IV (ThermoFisher), and 0.5 uL RNaseOUT (Invitrogen). After adding this RT mix to the RNA mix, I incubated the samples by the following gradient: 4 °C 2 min, 10 °C 2 min, 20 °C 2 min, 30 °C 2 min, 40 °C 2 min, 50 °C 2 min and 55 °C 10 min.

Oligo	Sequence
AA261	/5Phos/AGA/ideoxyU/CGGAA/ideoxyU/NNNNNNNNTTTTTTTTTTTTTTTTTTTTTTTTTTVN

Table 4.5 Oligo used for reverse transcription of polyadenylated RNAs prior to 3' PER barcoding

Following RT, I performed second-strand synthesis by adding 0.5 uL DNA Polymerase I, Large (Klenow) Fragment (NEB) and 0.1 uL RNaseH per 10 uL reaction, then incubating for 20 minutes at 12°C followed by 20 minutes at 22°C. I then incubated each sample with 1uL USER and 2 uL rSAP per 10 uL reaction for 15 minutes at 37°C and 5 minutes at 65°C. As with 5' PER, I then conducted the barcoding step according to the extension of single-stranded DNA in vitro method described above with the exception of using 10 uL of the RT reaction as the template and adjusting the amount of H2O accordingly. In the case of all 3' PER reactions, I used the inverted-dT hairpins depicted in Figure 4.1.B.

Library preparation

Tn5 cuts double-stranded DNA[114], therefore both 5' and 3' PER products at this point must be double-stranded. In the 3' PER protocol as described above, I use Klenow for second-strand synthesis. The 5' PER protocol however, requires me to amplify the library prior to tagmentation in order to generate double-stranded cDNA. Because the 5' PER protocol at this point is not a double-stranded library, I PCR amplified full-length cDNA prior to tagmentation. See Table 4.6 and 4.7 below for primer sequences used in the specified experiments.

Oligo	Sequence
AA223	AATGATACGGCGACCACCGAGATCTACACTAGATCGCTCGTCGGCAGCGTCAGATGTGTATAAGAGACAGAGTTTAAAG

AA225	CAAGCAGAAGACGGCATACGAGATTCGCCTTAGTCTCGTGGGCTCGGAGATGTGTATAAGAGACAG

Table 4.6 Oligos used in PCR amplification of full-length cDNA library generated with 5' PER in Figure 4.12

Oligo	Sequence
AA247	CCACTCACCTCCTTTAAACTGGGCCTTTTGGCCCAGTTTAAAGGAGGTGAGTGGGAGTGTTAGTTTTTTTT
AA248	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGAGTTTAAAGGAGGTGAGTGGGAGTGTTAG
AA249	AGTTTAAAGGAGGTGAGTGG
AA250	CCACTCCTTTAAACTGGGCCTTTTGGCCCAGTTTAAAGGAGTGGGAGTGTTAGTTTTTTTT
AA251	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGAGTTTAAAGGAGTGGGAGTGTTAG
AA252	AGTTTAAAGGAGTGG
AA253	ACATCACTTTAAACTGGGCCTTTTGGCCCAGTTTAAAGTGATGTGAGTGTTAGTTTTTTTT
AA254	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGAGTTTAAAGTGATGTGAGTGTTAG
AA255	AGTTTAAAGTGATGT
AA256	ACATCATTCCACTTTAAACTGGGCCTTTTGGCCCAGTTTAAAGTGGAATGATGTGAGTGTTAGTTTTTTTT
AA257	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGAGTTTAAAGTGGAATGATGTGAGTGTTAG
AA258	AGTTTAAAGTGGAATGATGT

Table 4.7 Oligos used in longer hairpin 4 design and subsequent PCR amplification of full-length cDNA library AA247, AA250, AA253, and AA256 are the same as listed above in Table 4.3, but are listed again here for convenience and association with their respective PCR primers. AA247 corresponds to +20 (long) hairpin v1, AA248 is the corresponding long primer and AA249 is the corresponding short primer. Following the same pattern of respective PCR primers, AA250 corresponds to +15 (short) hairpin v1, AA253 corresponds to +15 (short) hairpin v2, and AA256 corresponds to +20 (long) hairpin v2. I used these primers according to the samples tested in Figure 4.13 and 4.14.

From this point forward, library preparation is generally the same for both 5' and 3' PER. I cleaned samples according to manufacturer's instructions using AMPURE beads (Beckman Coulter) at 1.8X concentration, then performed tagmentation according to one of the following:

Tagmentation - Illumina TDE1

I added 8 uL 2x TD buffer at 1 uL of TDE1 to 7 uL of sample. I carried out tagmentation at 55°C for 5 minutes and stopped the reaction by adding 16 uL DNA binding buffer (Zymo) and incubating at room temperature for 5 minutes.

Tagmentation - Unloaded Tn5

For homemade and purchased unloaded Tn5, I loaded the oligos by first annealing the mosaic end reverse sequence to mosaic end adapter B (see Table 4.8 below) at a final concentration of 50 uM each in TE and incubating in the thermocycler for 3 minutes at 95°C, 3 minutes at 70°C, and colled at 2°C/minute from 70°C to 26°C. I diluted the

oligos 5x with H2O and added 1 volume of glycerol, prior to mixing the diluted oligos with Tn5 at a 4:1 oligo:Tn5 ratio and incubating for 30 minutes at room temperature. If not used immediately, I stored assembled Tn5 at -20°C.

Oligo	Sequence
AA242	/5Phos/CTGTCTCTTATACACATCT
AA244	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAG

Table 4.8 Oligos for loading Tn5 with adapter

To diluted 10 ng of sample in 9 uL, in the hood I added 4 uL TAPS buffer (50 mM TAPS-NaOH, 25 mM MgCl2 (pH 8.5 at room temperature)), 2 uL dimethylformamide, and 5 uL assembled Tn5. I mixed well by pipetting and incubated at 55°C for 5 minutes, then cooled the sample to 4-10°C. To stop the reaction, I added 4 uL of 0.1% SDS and incubated at 65°C for 10 minutes, then cooled to 4-10°C.

Following either tagmentation, I purified each sample according to manufacturer's instructions using AMPURE beads (Beckman Coulter) at 1.5X concentration and quantified the DNA with a Qubit 2.0 before PCR amplification. For final amplification of the 5' PER library, I used the same PCR primers used in the first amplification, AA223 and 225 (see Table 4.6). For 3' PER, I used AA225 and AA267 (see Table 4.9). I amplified each library according to Q5 (NEB) manufacturer's instructions for 20 cycles. Note that the melting temperatures and annealing temperature varies between the PCR primers used depending on the experiment.

Primer	Sequence
AA225	CAAGCAGAAGACGGCATACGAGATTCGCCTTAGTCTCGTGGGCTCGGAGATGTGTATAAGAGACAG
AA267	AATGATACGGCGACCACCGAGATCTACACTAGATCGCTCGTCGGCAGCGTCAGATGTGTATAAGAGACAGTCAAATTTCCTCCACTCACCCTCACAATC

Table 4.9 Oligos used for amplifying 3' PER libraries

After PCR, I cleaned each sample according to manufacturer's instructions with 1.8X AMPURE beads (Beckman Coulter) and used the Qubit 2.0 to obtain the library concentration.

Library quality, sequencing, and analysis

To assess library quality before sequencing, I ran each sample on an Agilent Bioanalyzer using the High Sensitivity DNA Kit. I sequenced samples on an Illumina MiniSeq using the Mid Output Kit (300-cycles). To assess read quality and obtain overrepresented reads, I used FastQC[116] through usegalaxy.org[117].

Results

Barcoded PER-poly dT hairpins efficiently extend single-stranded DNA *in vitro*

Barcoding in many split-pool single-cell RNA-sequencing methods is limited by the need to wash cells between each step, leading to sample loss. Because each round of PER-poly dT hairpin addition as shown in Figure 4.3 is specific to the round before, this would greatly reduce sample loss by eliminating washing at this stage of the experiment. Although this method is intended for use following reverse transcription, I decided to confirm that our design is efficient using a Cy5-modified single-stranded DNA template, as Kishi et al (2018) did in the original PER hairpin paper[115]. In full-scale split-pool barcoding experiments, each hairpin in a 96-well plate will have a different barcode, but for the purposes of the pilot experiments I describe here, I used the same set of hairpins for each experiment.

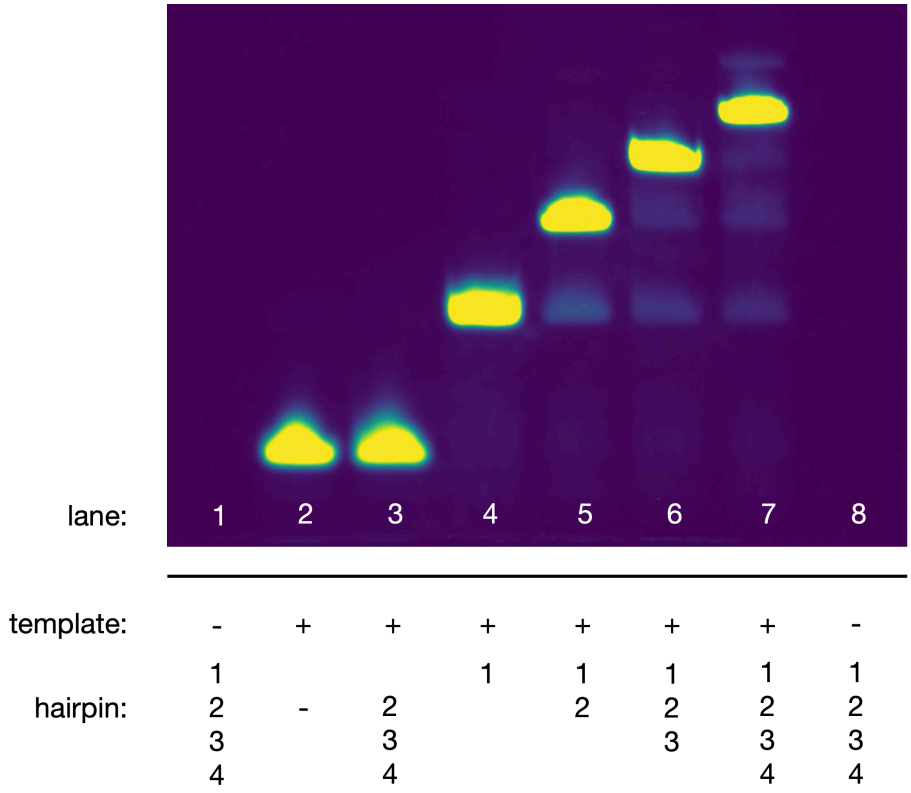

Figure 4.5 Efficient extension of single-stranded DNA *in vitro* using PER hairpins containing barcodes The template (absence or presence indicated by a – or +, respectively) is a short 14-mer modified with a fluorescent Cy5 molecule on the 5' end for visualization via TBE-PAGE. Each lane represents a single sample with the parameters below the gel in the figure. Lanes 1 and 8 represent the no template control in the presence of all 4 hairpins. Because the Cy5-modified oligo is not present in those lanes, no band is apparent. Lanes 2-7 do contain the Cy5-modified oligo in the presence of different combinations of hairpins as indicated above. Gel over-exposed to show faint bands, image is falsely colored on a scale of dark blue to yellow, and ladder not shown for figure simplification.

As shown in Figure 4.5 lanes 4-7, nearly all of the template is converted into the longest possible oligo based on the hairpins present during the reaction. I overexposed this gel (and all subsequent gels shown) when imaging to reveal any faint bands. The reactions are not 100% efficient due to

the presence of non-extended template below the brightest bands in lanes 4-7, but these results were extremely promising. The brightest bands being the longest bands in each lane also suggest that the PER reaction is highly specific and do not lead any duplicate additions of the same hairpin, which is a concern with ligation-barcoding and ineffective cell washing in some single-cell methods. Critically, in the absence of the first hairpin in lane 3, the template is not extended indicating that hairpin 1 is necessary for the extension by the following hairpins to occur. This, along with the lack of non-specific longer bands above the correctly barcoded (brightest) bands, indicate that the hairpins are specific to previous rounds of barcoding as expected. Ultimately, I have shown that not only are our PER hairpins extremely efficient, but also specific to the template or hairpins added.

Titrating hairpin concentration in the PER reaction reveals inefficiency following reverse transcription

Before attempting barcoding in a complex pool of RNA, I decided to optimize PER following reverse transcription with a purified 33-mer so that I could continue to conduct these experiments in the same manner as shown in Figure 4.5 and visualize every step of the reaction. I hypothesized that the presence of RNA and associated RT reaction components would negatively affect the PER reaction, but of course the RT step is absolutely necessary, so I conducted a series of experiments tweaking different parameters of the PER reaction itself.

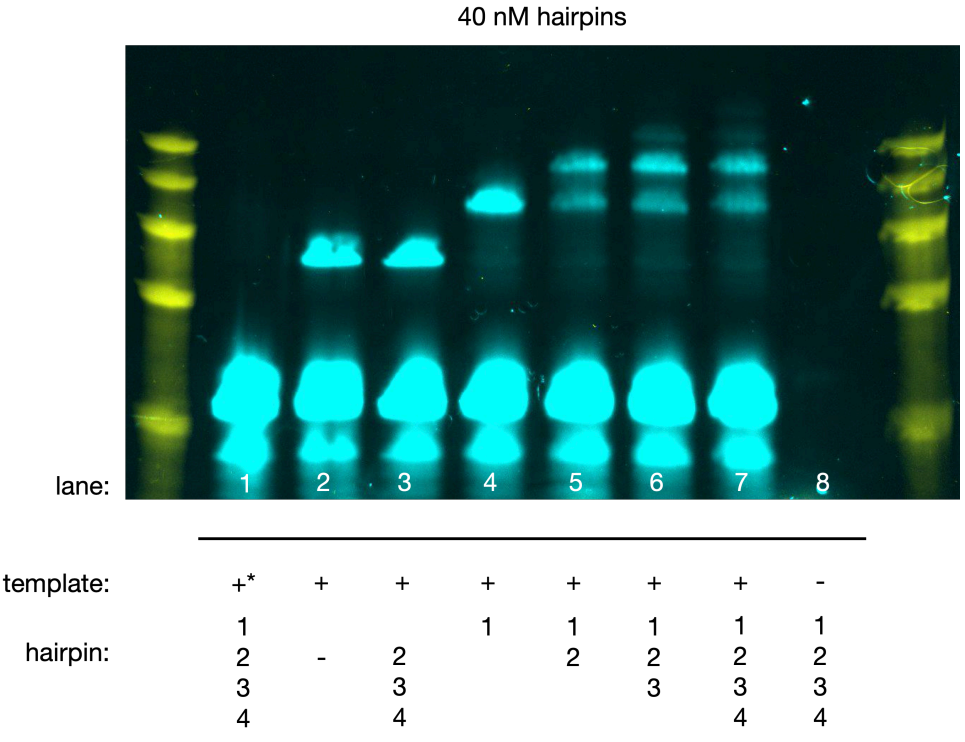

Figure 4.6 Reduced efficiency of PER extension following reverse transcription. The template is a 33-mer purified RNA oligo that is reverse transcribed with a Cy5-modified RT primer. In lane 1, the asterisk indicates that the template is present, but without a TSO as a control. Blue represents a Cy5 product and yellow is

the ladder. The shortest band in lanes 1-7 represents the RT primer and the band in each lane directly above that represents the reverse transcribed 33-mer. Gel over-exposed to show faint bands.

With the same reaction PER conditions as in Figure 4.5 along with the addition of reverse transcription and template switching, PER efficiency appears significantly reduced in the presence of more than just the first hairpin (see Figure 4.6). To me this indicates that the PER reaction is being limited by some amount whether a limiting reactant, presence of an inhibitor in the RT reaction, or some unknown reason. Additionally, the template-switching itself appears to be inefficient from the fact that the larger bright band (33-mer cDNA) is not efficiently converted to template-switched product, referencing the band above the template in lanes 2 and 3. With that being said, NEB (from whom I purchased the Template Switching RT Kit) reports a minimum template-switching efficiency of 20%.

I did not calculate exact efficiencies because I was never interested in those values, but I was not surprised by this observation and decided that inefficient template-switching was not an issue because the reaction works to some degree and I designed a template-switch oligo different from the manufacturer's suggestion for compatibility with PER. On the other hand, inefficient PER is an issue given the stark difference in efficiency before and after reverse transcription. In an effort to mitigate the reduction in efficiency, I decided to first test additional hairpin concentrations towards the maximum and minimum functional concentrations of PER hairpins.

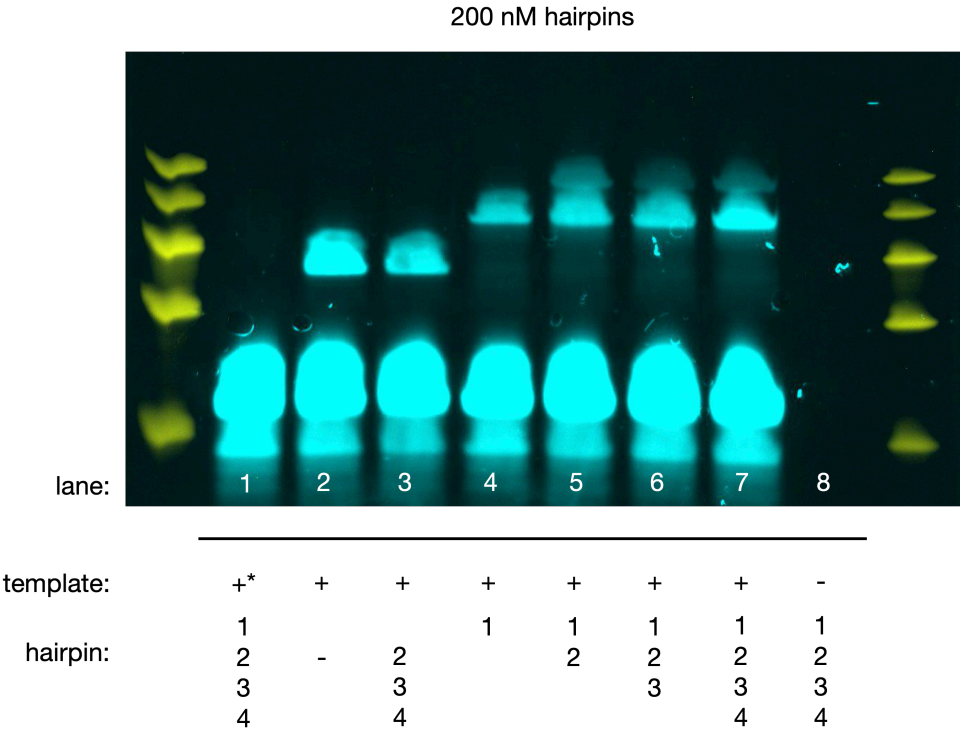

Figure 4.7 Increasing PER hairpin concentration further reduced efficiency of PER extension following reverse transcription. The setup for this experiment is the same as in Figure 4.8, the template is a 33-mer purified RNA oligo that is reverse transcribed with a Cy5-modified RT primer. In lane 1, the asterisk indicates that the template is present, but without a TSO as a control. Blue represents a Cy5 product and yellow is the ladder. The shortest band in lanes 1-7 represents the RT primer and the band in each lane directly

above that represents the reverse transcribed 33-mer. In this case however, I added a final concentration of 200 nM of each hairpin. Gel over-exposed to show faint bands.

As shown in Figure 4.7, increasing the concentration of hairpins to 200 nM appears to worsen the efficiency of the PER reaction relative to Figure 4.6, where I used 40 nM of each hairpin. Following this train of thought where more hairpins leads to a worse reaction efficiency, I decided to try a lower concentration of hairpins in case this would improve reaction efficiency.

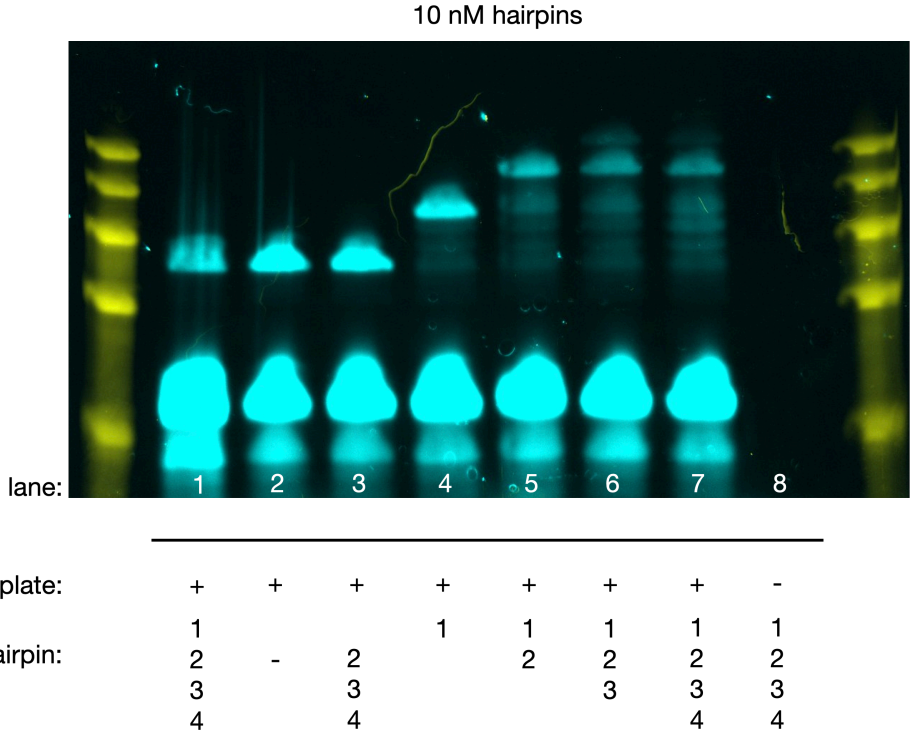

Figure 4.8 Decreasing PER hairpin concentration does not improve efficiency of PER following reverse transcription. The setup for this experiment is the same as in Figure 4.8 and 4.9, the template is a 33-mer purified RNA oligo that is reverse transcribed with a Cy5-modified RT primer. In this case, the TSO is present in the sample from lane 1. Blue represents a Cy5 product and yellow is the ladder. The shortest band in lanes 1-7 represents the RT primer and the band in each lane directly above that represents the reverse transcribed 33-mer. In this case, I added each hairpin to a final concentration of 40 nM. Gel over-exposed to show faint bands.

Similarly to Figure 4.6 where I used 40 nM hairpins, bands in the presence of the third and/or fourth hairpins above are hardly visible, but still present unlike Figure 4.7 where I used 200 nM hairpins. Because there were no stark differences in efficiency by eye between the 40 nM and 10 nM experiments, I decided to continue with a 40 nM hairpin concentration for all experiments moving forward. Following these experiments, I decided to try adding one hairpin at a time, mimicking the actual split-pool barcoding process, because I was previously incubating the template with all hairpins at the same time to make experiments easier. I also noticed that in each of the above experiments with only the first hairpin added during the PER reaction (lane 4), this particular reaction appeared to go to completion while subsequent addition of more hairpins did

not. From that observation, I posited that the problem is likely one of the reagents was limiting, something in the RT reaction was inhibiting PER, or an issue with molecular crowding.

PER incubation with one hairpin at a time indicates presence of limiting reagent or inhibitor

All PER reactions extending a single-stranded DNA oligo *in vitro*, and at least the first PER reaction following reverse transcription are efficient, but efficiency appears to taper off in subsequent rounds. Crowding seemed likely given the PER reaction is less efficient in the presence of more hairpin (Figure 4.7). In theory, if the problem were crowding and not limiting reagents or inhibitors, incubating each hairpin one at a time might mitigate any inefficiency by giving each PER hairpin more time with its respective template without crowding by subsequent addition of hairpins.

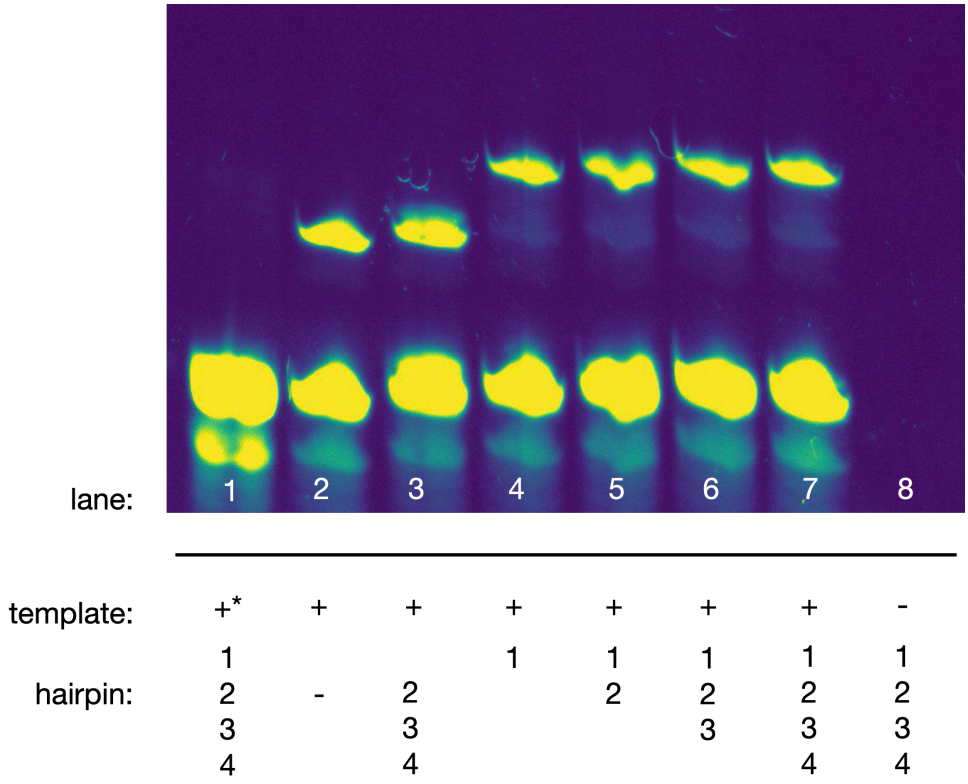

Figure 4.9 PER limited to the first step with sequential addition of hairpins following reverse transcription PER reactions in previous experiments occur simultaneously for four hours at 37°C, and the template is a 33-mer purified RNA oligo that is reverse transcribed with a Cy5-modified RT primer. In lane 1, the asterisk indicates that the template is present, but without a TSO as a control. The shortest band in lanes 1-7 represents the RT primer and the band in each lane directly above that represents the reverse transcribed 33-mer. In this case, I added hairpins sequentially starting with the first hairpin (unless otherwise indicated) and adding subsequent hairpins every hour, giving each hairpin 1 hour at 37°C with its template. Gel over-exposed to show faint bands, image is falsely colored on a scale of dark blue to yellow for visualization, ladder not shown for figure simplification.

As shown in Figure 4.9, PER reactions did not improve beyond the first round with sequential addition of hairpins. This suggests that the inefficiency seen in Figure 4.6, an experiment of the same design with the exception of sequential hairpin addition, is not due to molecular crowding. Together, the results of only the first round of PER occurring even in the presence of subsequent PER hairpins (Figure 4.9), or a reduced level of all four rounds together (Figure 4.6) indicate the presence of a limiting reagent or inhibitor in the reaction because of an overall reduction in successful extension of the single-stranded DNA template.

The PER reaction itself requires many reagents; however, considering the high efficiency of PER on its own and the addition of reverse transcription complicating the reactions in these cases, I hypothesized that the limiting reagent or inhibitor is a result of something preceding the PER reaction.

Supplementation of additional dHTP and longer inactivation of rSAP slightly improve PER efficiency following reverse transcription

The PER reaction relies on the absence of dGTP in solution for catalysis, yet dGTP are required for reverse transcription as RNA certainly contains all four dNTP. Given that no readily available method to rid the solution of a specific nucleotide exists, I eliminate all dNTP in these experiments enzymatically using Shrimp Alkaline Phosphatase (rSAP) before adding dHTP (dNTP with no dGTP) during PER. At the time, I was concerned about including too many reactions with high heat in the method, as eventually these reactions would be done in fixed cells or nuclei. With that in mind, I was heat inactivating rSAP at 65°C for only one minute to limit the amount of time the reactions were at higher temperatures.

Incomplete heat inactivation of rSAP would reduce the available dHTP during the PER reaction, as rSAP renders dNTPs unusable. At this point, I decided in addition to adding each PER hairpin sequentially, that I would try a more conservative approach of supplementing the PER reaction with additional dHTP before increasing the time the reaction spends at 65°C to inactivate rSAP.

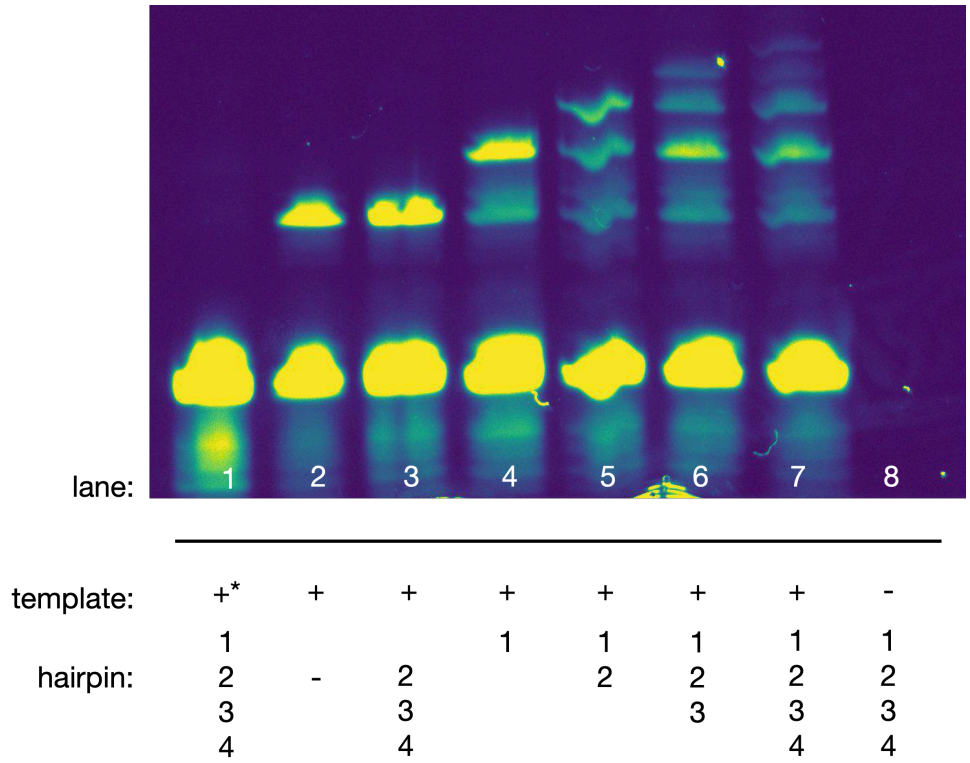

Figure 4.10 Supplementing dHTP after each round of PER improves reaction efficiency At the same time as adding the hairpins sequentially (one hour per round of PER), I supplemented the reaction with 5 uL of 1 mM dHTP the same amount added to the initial round. The template is a 33-mer purified RNA oligo that is reverse transcribed with a Cy5-modified RT primer. In lane 1, the asterisk indicates that the template is present, but without a TSO as a control. The shortest band in lanes 1-7 represents the RT primer and the band in each lane directly above that represents the reverse transcribed 33-mer. Gel over-exposed to show faint bands, image is falsely colored on a scale of dark blue to yellow for visualization, ladder not shown for figure simplification.

By adding 5 uL of 1 mM dHTP at the same time as the next PER hairpin, the efficiency of PER clearly improves in lanes 5-7 in Figure 4.10 relative to lanes 5-7 in Figure 4.9, as the bands representing addition of the 2nd, 3rd, and 4th hairpin are present. Without additional replicates and further analysis, I cannot say whether or not efficiency is different between the experiment represented by Figure 4.6 and Figure 4.10. However, a clear improvement occurs between the experiment with sequential addition of hairpins (Figure 4.9) and sequential addition of hairpins coupled with supplemental dHTP (Figure 4.10).

Even with the promising improvement from this more conservative approach, not changing the time spent inactivating rSAP at 65°C and thinking ahead to how cells and nuclei would respond at higher temperatures, I still wanted to test a longer heat inactivation. If rSAP persistence is ultimately the issue, supplementing the reaction with dHTP would not exactly solve the problem. According to the manufacturer (NEB), rSAP is nearly entirely deactivated after 5 minutes at 65°C, so I conducted the same experiment without supplemental dHTP, but with longer rSAP heat inactivation.

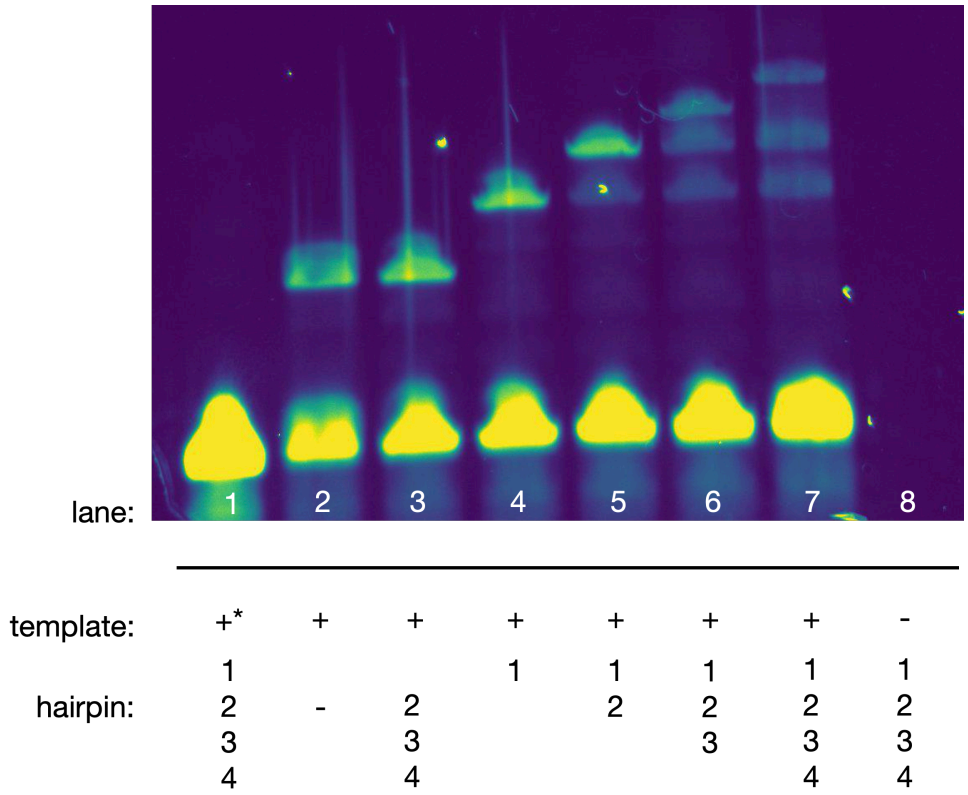

Figure 4.11 5 minute rSAP heat prior to PER improves reaction efficiency The template is a 33-mer purified RNA oligo that is reverse transcribed with a Cy5-modified RT primer. In lane 1, the asterisk indicates that the template is present, but without a TSO as a control. Blue represents a Cy5 product and yellow is the ladder. The shortest band in lanes 1-7 represents the RT primer and the band in each lane directly above that represents the reverse transcribed 33-mer. In this case, each hairpin was added sequentially and no additional adjustments were made during PER. Gel over-exposed to show faint bands, image is falsely colored on a scale of dark blue to yellow for visualization, ladder not shown for figure simplification.

Relative to the experiment with the only adjustment being sequential addition of hairpins (Figure 4.9), PER is more efficient with sequential addition of hairpins coupled with a longer (5 minute) rSAP heat inactivation. With that considered, the efficiency of PER following reverse transcription still does not reach that of PER on a single-stranded DNA template. I also conducted experiments supplementing the reaction with additional Bst polymerase, as well as looked for restriction of PER on the single-stranded DNA template in the presence of reagents used during reverse transcription.

I did not obtain any significant results from these experiments, so I decided to move forward testing reverse transcription followed by PER in total RNA extracted from embryos, even though PER efficiency was less than ideal. Ultimately, I wanted to know if this is a viable method for barcoding RNA and to some extent, the product of final round of PER in the experiments above is detectable (Figure 4.6, Figure 4.10, Figure 4.11). Barcoding efficiency is will need to be revisited *in vitro*; however, next-generation sequencing provides millions of reads, which is more than enough to determine whether or not barcoding is at least partially successful in total RNA.

Pilot sequencing of PER-barcoded total RNA suggests inefficient PCR amplification

As shown in Figure 4.4.A, 5' PER barcoded RNA must be PCR amplified to make all products double-stranded prior to tagmentation and final amplification. Rather than continuing with tagmentation, which requires expensive reagents, I decided to sequence the cDNA prior to tagmentation. Before sequencing a full library, I assess library size and quality using an Agilent Bioanalyzer. Sequencing untagmented cDNA poses its own challenges in that most sequencing methods require average library size under 1000 bp and the average *Drosophila melanogaster* cDNA is much larger; however, there should be enough shorter barcoded cDNAs present that I am able to sequence and I would be able to observe the overall size distribution from the Bioanalyzer traces.

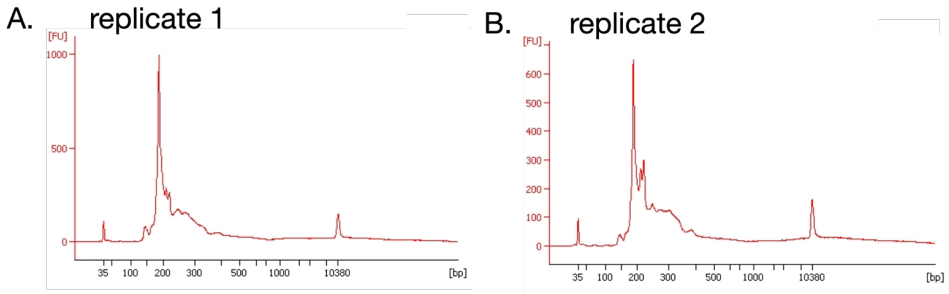

Figure 4.12 Pilot cDNA library does not correspond to expected size Figure depicts Bioanalyzer traces for two replicates (A and B) of full-length cDNA library generation by first reverse transcribing total RNA from embryos using a random hexamer primer with a PCR handle, followed by 5' PER barcoding and PCR. I would expect a successful cDNA library generation using random hexamers to follow a normal distribution somewhere around 2000 bp, but skewed some amount shorter given the random hexamers are going to bind randomly and not necessarily produce a full-length cDNA.

Because I used a random-hexamer RT primer, I expected the distribution of the library to fall along the distribution of RNA lengths likely skewed shorter than actual given an RT primer is certainly not going to bind at the very end of the transcript each time. However, as shown in Figure 4.12, whatever products present following PCR amplification on average were much short with sharp peaks around 200 bp. Because the primers (79 bases) and hairpins (59 bases) are long, I assumed that these short sharp peaks were some sort of dimers of either of these reagents, as this is a common issue in sequencing library preparation. The signal at these shorter lengths is so high, even if PER barcoding worked, the sample is overwhelmed by the wrong product. I proceeded with sequencing libraries anyways to be certain.

Indeed, the vast majority of the reads were not what I would expect for successful PER barcoding (data not shown). Unfortunately the beginning of the reads also did not correspond to bases at the end of the primer, which I would expect if I was randomly sequencing other cDNAs that were not necessarily barcoded, but by chance happened to have the same sequence as the end of the PCR primer. Because I expected some cDNAs to be amplified due to matching the PCR primer by chance and I did not observe any, I presumed that the PCR itself was doomed to fail.

Longer universal PER extension improves PCR amplification

In a hypothetical single-cell RNA-sequencing experiment using PER where each hairpin is uniquely barcoded, the maximum number of bases in common to PCR amplify from would be 9 bases. At the time, I knew a PCR primer with such a short overhang had a high chance of failure. In order to overcome the overhang problem, I designed multiple 4th round PER hairpins with the barcode removed, but would add a longer sequence to amplify from. According to NEB, the maximum length of strand displacement activity, which is required for PER, by Bst Polymerase is 20 base pairs, thus I limited the added length to 20 bases (see Table 4.3 for oligo sequences). I designed multiple versions of longer PER hairpins in order to test for increased PCR efficiency, but first I wanted to ensure that these new hairpins performed as expected *in vitro*.

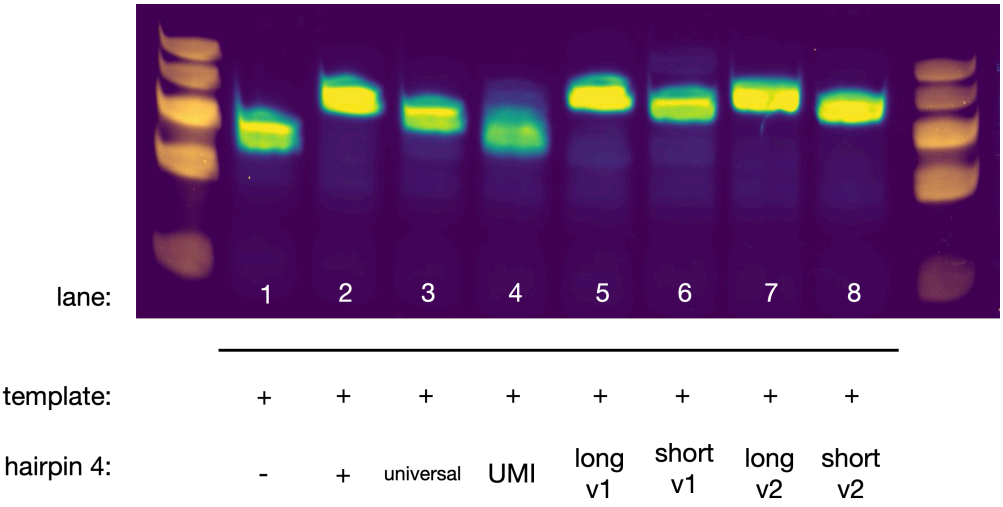

Figure 4.13 Successful PER occurs with longer landing regions in the 4th round The template (is a short 14-mer modified with a fluorescent Cy5 molecule on the 5' end for visualization via TBE-PAGE. In this case, I performed PER without reverse transcription to simplify the longer 4th round PER hairpin test. Lane 1 represents PER extension through the 3rd round in the absence of any 4th round hairpin. Lane 2 represents PER extension through the 4th round with the same 4th hairpin used above containing the 3rd hairpin landing site and 5 base barcode, adding a total of 14 bases to the template. Lanes 3-8 represent several new designed 4th round hairpins. Lane 3 hairpin 4 is a universal version of the hairpin in lane 2, universal meaning without the 5 base barcode. Lane 4 hairpin 4 is a universal hairpin designed to also add a unique molecular identifier (UMI). Lanes 5-8 hairpin 4s designed to add either long (20 base) or short (15 base) universal sequences. Different versions of lanes 5-8 hairpin 4s indicated with v1 (version 1) or v2 (version 2). Gel over-exposed to show faint bands, Cy5 channel is falsely colored on a scale of dark blue to yellow.

Distinguishing 14 base from 15 and 20 base bands on the TBE-PAGE gel in Figure 4.13 is difficult; however, the difference between 5 and 7 relative to 6 and 8 is clear and indicates that in these cases, the universal 4th round of PER is successful with a longer extension. In this experiment, I decided to also test the addition of a Unique Molecular Identifier (UMI) with PER as shown in Lane 4. UMIs are critical for the proper quantification of gene expression[118,119], but I had not yet considered at what step of the protocol when I would add UMIs. Based on the experiment shown in Figure 4.13, I will have to find an alternative solution for adding a UMI as

the band appears the same length of the Lane 1 control indicating almost entirely unsuccessful PER.

A universal 4th PER hairpin would limit the number of single cells that could be sequenced by limiting the possible barcode combinations; however, this was a critical problem to solve. In the future in order accommodate a larger sample size (greater than roughly 50k cells) I could design a similar extended hairpin for a 5th round of PER so that I can still accomplish 4 total rounds of barcoding. After establishing that these universal 4th PER reactions were successful, I needed to determine if the sequences added would actually improve full-length cDNA amplification.

Following the experiment shown in Figure 4.13, I conducted an experiment using the long v1 and v2 hairpins that add 20 universal bases in the 4th round of PER to determine whether or not the longer PCR primer overhangs lead to an improvement in cDNA amplification. To control for the long primer, I also designed a short PCR primer (20 bases) that entirely overlap with the sequence added by the hairpins. In this control, any PCR amplification issue should not result from overhang issues.

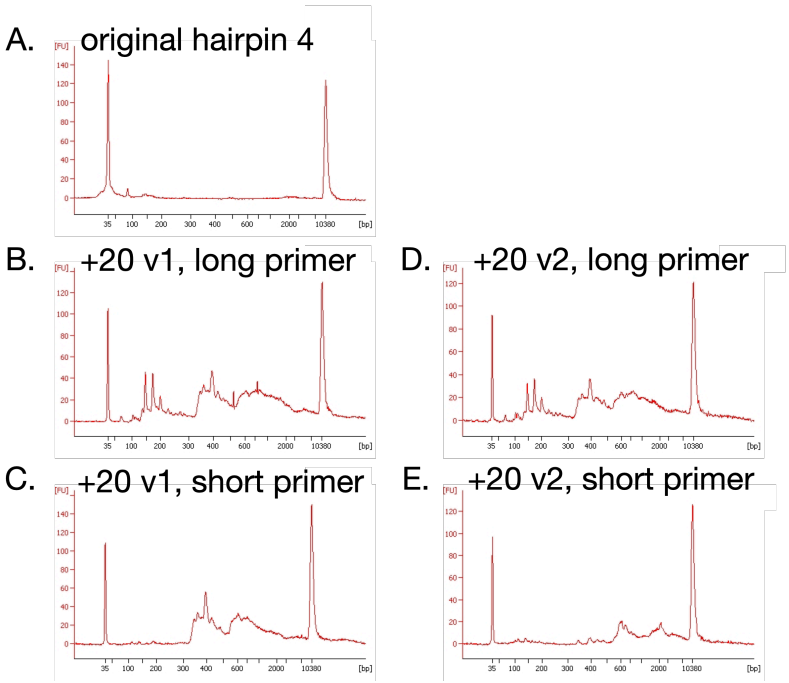

Figure 4.13 Addition of a longer universal sequence in the 4th round of PER appears yields a potential cDNA library following PCR with long and short primers Figure depicts Bioanalyzer traces for (A) the original hairpin 4 design amplified with a primer with a short (9 bp) overhang, (B) a newly designed hairpin 4 that adds 20 bases and is amplified with a long PCR primer and (C) short PCR primer, and (D) a second version of a designed hairpin 4 that adds 20 bases and is amplified with a long PCR primer and (E) short PCR primer.

This time, amplification of anything with the original hairpin 4 with a 9 base PCR primer overhang failed entirely (Figure 4.13.A), thus I was unable to replicate previous experiments (Figure 4.12). However, I was able to amplify cDNA to some extent from both universal hairpin

4 designs with a long primer (20 base overlap) and short primer (20 base total) (see Figure 4.13 B-E). One cannot read too far into Bioanalyzer traces as the composition of a library will not be apparent until sequencing, but I arbitrarily chose to continue with library preparation for the full-length cDNA coupled with universal 4th round PER by the long hairpin v1.

When first designing the theoretical PER snRNA-seq method, I thought that priming reverse transcription with random hexamers would provide a better representation of nuclear data, as many RNAs may not be polyadenylated. By conducting PER with a single hairpin for each barcoding round, in this case 1-3, all of the reads corresponding to this side of each molecule should be the exact same and I could easily detect them by observing the sequences that appear the most. However, in the first pilot sequencing experiment using random hexamer RT primers), I noticed that I was essentially sequencing ribosomal RNA (rRNA) repeatedly and not capturing any PER barcoded reads by running the overrepresented sequences in BLAST[120].

Ribosomal RNAs are highly abundant in eukaryotic cells, approaching 80% in fast-developing mammalian cells for example[121], but I presumed that successful PER might overwhelm signal from rRNAs; however, this turned out to not be the case. I was aware that this might be an issue as many RNA-sequencing experiments rely on polyA+ selection or rRNA depletion to ensure that rRNAs, which comprise the vast majority of RNAs present in eukaryotic cells, do not constitute most of the sequencing reads[122]. This proved to be the case in my experiments, but to my knowledge polyA+ selection and rRNA depletion have never been done *in situ*.

As an alternative, I decided to pursue priming reverse transcription with poly-dT primers to only capture polyadenylated RNAs. Although this would exclude non-coding (important, but non-polyadenylated RNAs), capturing only polyadenylated RNAs would certainly still yield interesting results as these RNAs represent the molecules that are translated into proteins, and using poly-dT primers easily solves the rRNA problem.

Poly-dT cDNA libraries correspond to expected sizes of full-length cDNA, but are not successfully barcoded

Priming RT with a PCR primer handle ending in a string of random hexamers would allow me to capture non-coding RNAs and other RNAs that are non-polyadenylated; however, in that case rRNAs will comprise the bulk of sequencing data. Given that I will not be able to deplete rRNAs *in situ*, I decided to start priming RT with a poly-dT instead. I conducted the following experiment in the same manner as in Figure 4.16, with the RT primer being the only difference. Additionally, I expected the size distributions of RT primed with a poly-dT to follow a distribution around the average cDNA size, or 2 kb[123].

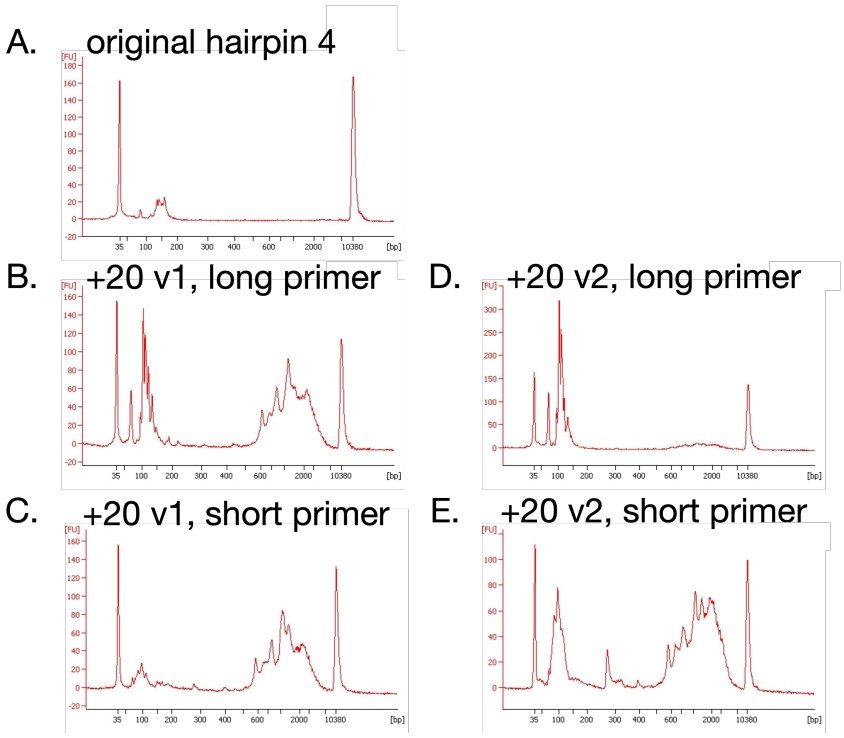

Figure 4.14 Priming RT with poly-dT followed by PER yields a full-length cDNA library at the expected size Figure depicts the same experiment as 4.16 with the exception in each of these cases being that RT was primed with a poly-dT primer instead of random hexamers. Bioanalyzer traces for (A) the original hairpin 4 design amplified with a primer with a short (9 bp) overhang, (B) a newly designed hairpin 4 that adds 20 bases and is amplified with a long PCR primer and (C) short PCR primer, and (D) a second version of a designed hairpin 4 that adds 20 bases and is amplified with a long PCR primer and (E) short PCR primer.

Again, when conducting PER with the original hairpin 4 and short overhang PCR primer, there is a peak at low sizes, but none of the expected cDNA. However, the broad peaks around 2 kb in Figure 4.14.B, C, and E were promising as they correspond to the expected cDNA library size. A very small, broad peak is apparent following PER with the second version of the extended universal hairpin; however, PCR amplification appeared inefficient in that case (Figure 4.14.D). As such, I decided to continue with the first version of this new hairpin 4 in subsequent experiments.

PCR amplification clearly occurs with the long and short primer (Figure 4.14.B and C), but the library amplified with the long primer contains a greater amount of short products, unlikely to be cDNA and likely to be primer dimers. With that being said, long overlapping primers are necessary to continue with library preparation in order to add sequencing adapters (see Figure 4.4). Fortunately, I can eliminate dimers and other unwanted short products with size selection in order to prevent their amplification in subsequent reactions. At this point, I decided to continue with library preparation with the sample shown in Figure 4.14.B.

Upon looking at the overrepresented sequences in the library following sequencing, I realized that the hairpins themselves were being repeatedly sequenced in some capacity as I noticed poly(8)-dT stretches in these sequences. The poly(8)-dT sequence at the end of each hairpin

serves to prevent the polymerase from extending the hairpins themselves[115], but I could not figure out what exactly happen molecularly whether hairpin extension or something else. Fortunately Kishi et al 2018 also conducted experiments using a single inverted-dT to prevent hairpin extension[115], so I decided to try that instead.

Barcoded PER-inverted-dT hairpins efficiently extend single-stranded DNA *in vitro*

Before using the inverted-dT hairpins to test PER following reverse transcription, I first wanted to replicate the efficiency observed by the poly-dT hairpins *in vitro* so I conducted the same experiment as described previously with the same conditions, but with these new hairpins instead (see Figure 4.5).

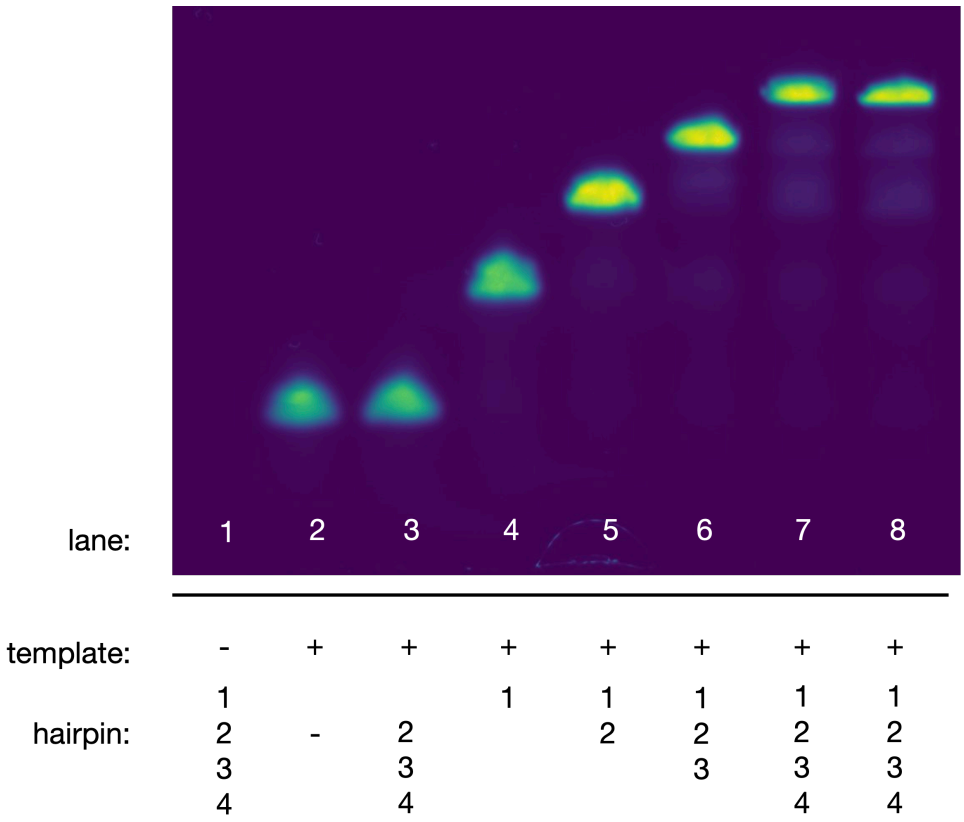

Figure 4.15 Redesigned PER hairpins ending in a single inverted-dT efficiently extend single-stranded DNA *in vitro* The template (absence or presence indicated by a – or +, respectively) is a short 14-mer modified with a fluorescent Cy5 molecule on the 5' end for visualization via TBE-PAGE. Each lane represents a single sample with the parameters below the gel in the figure. Lanes 1 and 8 represent the no template control in the presence of all 4 hairpins. Because the Cy5-modified oligo is not present in those lanes, no band is apparent. Lanes 2-7 do contain the Cy5-modified oligo in the presence of different combinations of hairpins as indicated above. Gel overexposed to show faint bands, image is falsely colored on a scale of dark blue to yellow, and ladder not shown for figure simplification.

The efficiency of single-stranded DNA extension by the redesigned PER hairpins with an inverted-dT parallels that of PER hairpins with a poly-dT tail (see Figure 4.18 and Figure 4.7). Faint bands are hardly visible in lanes 6-8 below the correctly barcoded products represented by the brightest bands, even after overexposing the gel. Although faint bands indicate incompletion, the fact that these bands are only visible following overexposure is promising.

Both iterations of PER hairpin design are highly efficient *in vitro,* but because the poly-dT hairpins are somehow incorporated into the final product, I decided to continue with just the inverted-dT version of the hairpin. Because the hairpins are also amenable to any method that requires extension of single-stranded DNA, I also decided to pursue 3' end sequencing rather than generating full-length libraries then 5' end sequencing as described above.

Pilot 3' PER libraries did not yield expected barcoded product

In theory, sequencing the 3' end of cDNA would reduce the number of required PCR amplifications from two to one based on our design (Figure 4.4). PCR amplification bias is known to affect library preparation and sequencing[124]. While I could optimize the reaction, eliminating one PCR reaction entirely is guaranteed to reduce bias from PCR amplification. Before optimizing each step of the reaction similarly to Figures 4.6-4.11 above, I conducted a pilot experiment in its entirety first to determine if the method works in some capacity as is.

At the same time, I tested three different Tn5 enzyme preparations for tagmentation because this is a necessary step that I had not yet addressed. Illumina TDE1, their version of Tn5, comes pre-assembled with two different sequencing adapters, whereas I would to load the adapters for homemade Tn5 myself. With either 5' or 3' PER, I will be sequencing a particular end of the cDNA library, but I would lose half of the potential sequencing reads due to PCR incompatibilities using Illumina TDE1. Using homemade Tn5, I am able to load Tn5 with one adapter to not lose any reads. This option is more cost-effective in terms of making the enzyme itself and not losing sequencing reads due to incompatible adapter-primer pairs, but homemade Tn5 does not come with the quality assurance of pre-loaded proprietary Tn5. In the future, if people using this or other single-cell methods have the option to use homemade Tn5, that would be preferable to Illumina TDE1; however, either should work in the end.

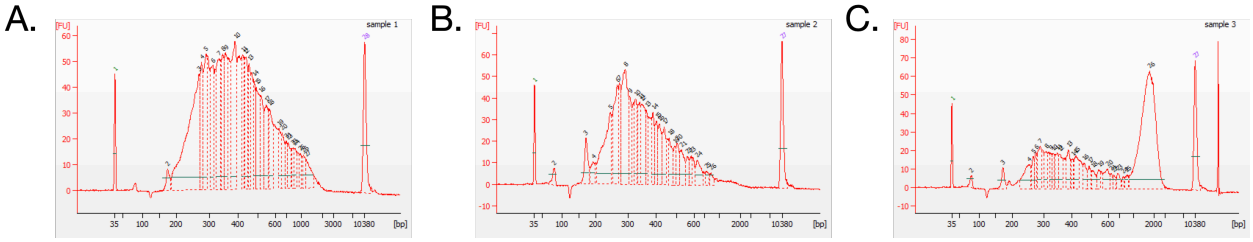

Figure 4.16 Tagmentation following pilot 3' PER barcoding yields sequenceable library The Bioanalyzer traces above represent 3' PER libraries generated using (A) Illumina TDE1, (B) Tn5-George, and (C) Tn5-MacroLab. Illumina TDE1 (A) and Tn5-George (B) libraries appear as broad peaks averaging around 400 bp, indicating successful tagmentation occurred. Numbers and dashed lines indicate detected peaks. The large spike averaging at 2000 bp in the Tn5-MacroLab (C) library indicates incomplete tagmentation.

Following 3' PER barcoding and presumably successful tagmentation, I would expect PCR amplification to yield a library averaging around 400 bp. This appears to be the case when performing tagmentation with TDE1 from Illumina as well as unloaded Tn5 kindly prepared by a colleague in Don Rio's lab, George Ghanim and my own colleague (and best friend) Jenna Haines. However when using the Tn5 prepared by the Berkeley MacroLab, a large peak remains at 2000 bp. This suggests incomplete tagmentation, as not all of the barcoded cDNA was cut into shorter fragments. However, this is unexpected as untagmented product should not be amplified because Tn5 would not have deposited the necessary adapters for subsequent PCR amplification. There is no way to know the content in a library without sequencing. However, if the 2000 bp peak in Figure 4.16.C is truly the untagmented product, that indicates that the PCR primers might possibly correspond to *Drosophila melanogaster* cDNA sequences by chance. With that being said, I decided to sequence the 3' PER library tagmented with Illumina Tn5 because this library appears sufficiently tagmented and I wanted to determine if barcoding worked to any extent.

Unfortunately, my hypothesis that the PCR primers were amplifying *Drosophila* cDNA by chance was correct. The 5' end of most overrepresented sequences as determined by FastQC[116] correspond to the end of the respective PCR primer for sequencing read 1 and read 2. I did not detect any correctly barcoded sequences within these data, but that does not mean that 3' PER sequencing will not work. This was a pilot experiment, and further experiments are necessary to optimize each step.

Discussion

Here, I demonstrate the possibility of using catalytic hairpins to conduct split-pool barcoding in single-cell RNA-sequencing. From this study I found that our barcoded PER hairpins based on Kishi et al 2018[115] efficiently extend single-stranded cDNA *in vitro* (see Figure 4.5 and Figure 4.15); however, further work needs to be done to optimize the PER reaction in a more complex mixture. The *in vitro* reaction mixture is simple, but a PER reaction following reverse transcription requires many additional steps that may explain why the reaction is not as efficient in this case.

The reaction is consistently efficient when extending a Cy5-modified single-stranded DNA template between hairpins preventing extension with a poly(8)-dT or inverted-dT (Figures 4.5 and 4.15 respectively). However, under the same conditions PER following reverse transcription is clearly not as efficient (Figure 4.6). No matter if I alter the concentration of hairpin (Figures 4.6-4.8), add the hairpins one at a time (Figure 4.9), supplement the reaction with dHTPs (Figure 4.10), or increase heat inactivation time of rSAP (Figure 4.11), the reaction still does not approach the efficiency seen in the PER reaction without RT. These modifications were tested with 5' PER and should be repeated with 3' PER as well to determine if 3' PER works to some extent *in vitro*.

One consideration that I was unable to address within the scope of this dissertation, is that RNA quality may be affected by fixation. RNA quality is important for 3' PER as well; however, 5' PER relies on the generation of a full-length cDNA library. Obtaining quality RNA after fixation

is certainly possible in general[125–127], and has been done for single-cell RNA-sequencing as well[128], but this is something that must be considered for 5' PER in the future. I presume that non-full-length 3' PER libraries may reduce fixation bias because extension must span a few hundred bases rather than the entire transcript, nonetheless we should study and utilize any strategy that might mitigate sequencing bias resulting from any part of the protocol.
Although I was unable to replicate the efficiency of the initial PER experiments for PER following RT *in vitro*, I did observe some amount of fully barcoded cDNA under certain conditions (see Figures 4.6, 4.8, 4.10, and 4.11). PER barcoding addresses limitations of existing barcoding strategies by relying on barcoded oligos as opposed to an expensive proprietary enzyme and reducing sample loss by eliminating washing steps. Given the difficulty of isolating cells or nuclei from certain samples, especially in perturbation experiments, the reduced sample loss from not having to wash the samples in between each barcoding round likely outranks the disadvantage of a not completely efficient PER. Current sequencing technologies now yield over a billion reads, so even if one has to discard a large portion of reads due to some inefficiency, one can always sequence more of the library, but not necessarily obtain more of the original sample. With further optimization, PER barcoding will allow us to sequence tens of thousands to potentially millions of single cells or nuclei with reduced sample loss and reduced reliance on proprietary reagents.

Chapter 5: Concluding Thoughts and Future Directions

For well over 100 years, researchers have used *Drosophila melanogaster* as a model system to understand genes and genomes. Since then, we have learned much about how gene expression is regulated and how genomes are structured. Enhancers regulate gene expression across space and time via transcription factor binding, accessible chromatin, and potentially association with chromatin modifications. Other genomic elements, such as insulators, are associated with enhancer function as well as chromatin structure. During early development in particular, the regulation of gene expression and genome structure is driven by maternally-deposited RNAs and proteins prior to activation of an organism's own genome. As discussed in Chapter 1, despite the massive body of knowledge that scientists have accrued over all this time, our understanding of the establishment of gene regulation and genome structure remains incomplete.

Enhancer activity is largely driven by coordinated transcription factor binding[8,9]; however, conservation of patterned gene expression despite sequence divergence suggests that transcription factor binding does not entirely define enhancer identity[24]. Other features, such as chromatin accessibility[31,32,34], chromatin modifications[38,41], and insulators[43,44] are associated with enhancer function as well. In Chapter 2, I showed that expression of a patterned gene is affected upon knockdown of multiple chromatin modifiers, while expression of a ubiquitously-expressed gene as a proxy for transcription in general remains normal. Loss of these chromatin modifiers also results in reduced embryonic viability. These results suggest that some chromatin modifiers may specifically participate in the regulation of enhancer activity or the specification of enhancer identity. However, when confirming the extent of knockdown in one of the candidates, I found a wide and overlapping variability in its expression between control and knockdown conditions.

From the above results, I concluded that further experiments should utilize a more consistent method of perturbation, such as germline clones. Also, at the time my intentions were to conduct further experiments using single-embryo RNA-sequencing which would allow me to examine overall gene expression; however, I would not be able to observe any patterning changes which may indicate an effect on enhancer activity. As early development is largely controlled by maternal RNAs and proteins[63], any single-embryo RNA sequencing results also would certainly not reflect zygotic gene expression levels. Fortuitously, single-cell and single-nucleus RNA-sequencing technologies slowly improved and became increasingly accessible throughout my PhD. Single-nucleus RNA-sequencing in the early *Drosophila* embryo would allow for the study of both gene expression level upon loss or mutation of important developmental regulators and effects on spatially-patterned gene expression.

In the middle of my PhD I became increasingly interested in the role insulators have on enhancer-driven gene expression, following a publication from another member of the lab[50]. Interestingly, an insulator protein, CTCF, is essential for survival as the only insulator protein in mammals. *Drosophila melanogaster* actually have several insulator proteins however[46,47], allowing us to study the function of insulator proteins without evoking cell death. I did investigate *Drosophila* CTCF (dCTCF) in the screen described in Chapter 2; however, I did not observe any interesting phenotypes. Whether this is biological, or a result of potential variability in knockdown as discussed above is unknown. Nonetheless, I decided to combine my newfound

interest in insulator function with my interest in single-nucleus RNA-sequencing as a means to understand the regulation of patterned gene expression in the early embryo.

Despite previous reports of minimal change in overall gene expression, loss of dCTCF has been shown to result in changes in patterned gene expression[54,55]. Given this, along with my thoughts on how single-embryo RNA-sequencing, or bulk sequencing in general, provide no patterning information, I believed that single-nucleus RNA-sequencing would be a better method to find genes with altered gene expression. In Chapter 3, I describe a single-nucleus RNA-sequencing experiment and series of analyses I conducted in pre-cellularization *Drosophila melanogaster* control and *dCTCF*$^{mat-/-}$ embryos. After adequate filtering of the data and ensuring the data are of high quality, I showed that this technique allows for the detection of known nucleus types in the embryo. Although I removed yolk and pole cell nuclei, as these are not necessarily interesting in looking at developmental patterned gene expression, their presence provided assurance that the technique works and the data are sound. Prior to conducting single-nucleus RNA-sequencing, I was not sure whether or not the gene expression data in individual nuclei would retain spatial information. I found that the nuclei actually do cluster according to spatial regions of the embryo, which allowed me to then ask questions regarding differential gene expression across the embryo.

Although I cannot relate the exact number of differentially expressed genes to previous papers, as analyses and conditions may differ, I did find many cases of differential expression unlike previous reports. In addition, I found many examples of spatially patterned genes that are not differentially expressed in bulk (across all nuclei) but are in specific clusters. Altogether these results demonstrate that single-nucleus RNA-sequencing can be used to understand the regulation of patterned gene expression. In order to understand exactly how these changes in expression affect embryonic development and patterning, I would move forward by conducting single molecule RNA FISH (smFISH) of nuclear RNAs to quantify changes in zygotic transcription. These experiments are outside of the scope of this dissertation; however, the analysis I have conducted opens up many doors for exploration. I believe this method will be valuable in understanding changes in gene expression upon the loss of any factor playing a role in early development that relates to transcription. In addition to establishing single-nucleus RNA-sequencing in the early *Drosophila melanogaster* embryo, I made all of my analyses publicly available in notebooks that can be run directly online so that others can explore expression of genes that I may not have looked at. I am proud of what I was able to do with this, despite having no prior experience in bioinformatics, and making my code available is my way of paying it forward.

Open access availability and usability of papers, methods, and code is one of my core values as a scientist. Despite my success with a commercially available single-nucleus RNA-sequencing method, I feel that the field has a long way to go to lower the cost and increase the access of these technologies. As an example, the early *Drosophila* embryo contains on the order of 5,000 to 6,000 nuclei, but the single-nucleus technology I used in Chapter 3 captures up to 10,000 cells or nuclei. In order to more definitively answer questions about biology rather than generating a list of candidate genes that are changed as described above, I would need to sequence many more replicates. However, the cost of just a single experiment precluded me from doing so. In an effort to drive the single-cell and nucleus sequencing field forward, I began a collaboration with Lior

Pacther's group at CalTech. As described in Chapter 4, we have designed and I have tested a set of catalytic DNA hairpin oligos in an effort to reduce the cost of barcoding single cells as well as reducing sample loss. At its core, barcoding is simply a specific extension of DNA molecules. To do so, we added barcodes to an existing catalytic DNA hairpin oligo concept[115].

I was able to reproduce the efficiency seen from the original work with our barcoded hairpin design *in vitro* as well as the specificity of each round of extension. This is important to note, as some existing single-cell barcoding methods rely on ligation. Ligation reactions are highly efficient; however, require for cells (or nuclei) to be washed in between barcoding rounds. Washing in general results in sample loss, and incomplete washing can result in multiple cells having the same barcode, neither of which are good in single-cell experiments. Despite high efficiency of single-stranded DNA *in vitro* however, I faced several challenges in optimizing this reaction following reverse transcription, a necessary step in developing this method for single-cell RNA-sequencing. I tried multiple strategies that mildly improved barcoding efficiency following reverse transcription; however, none of these strategies increased the reaction efficiency enough to match the efficiency of extension of single-stranded DNA shown in the initial experiments. At the time, I decided to continue optimizing downstream reactions in the experiments to establish a draft version of the method as a whole. I found library preparation issues in an attempt to capture one end of the cDNA molecule, however I believe with further optimization we will be able to capture either end of a cDNA molecule. This collaboration is ongoing and will not be completed at the end of my PhD; however, I believe I have made significant progress towards thinking about and improving access to single-cell RNA sequencing to answer interesting biological questions.

Altogether, the work I have done throughout my PhD has provided me with many skills and informed how I think about science in general. Sometimes we need to think beyond conventional ideas, like features other than transcription factor binding sites contributing to enhancer identity. Sometimes we have to go out of our comfort zone and try something that has never been done before to move things forward, like with single-nucleus RNA-sequencing in pre-cellularization embryos. Science is about working together and sharing what we learn in order to generate progress, as with the collaboration that has formed during my PhD and will continue beyond. Most importantly, I have learned that you can be a scientist no matter where you start your journey and that we all have the power to make science a better place for everyone.

References

1. Morgan, T. H., Sturtevant, A. H., Muller, H. J. & Bridges, C. B. *The Mechanism of Mendelian Heredity*. (Holt, 1922).
2. Franklin, R. E. & Gosling, R. G. Molecular configuration in sodium thymonucleate. *Nature* **171**, 740–741 (1953).
3. Banerji, J., Rusconi, S. & Schaffner, W. Expression of a beta-globin gene is enhanced by remote SV40 DNA sequences. *Cell* **27**, 299–308 (1981).
4. Nüsslein-Volhard, C. & Wieschaus, E. Mutations affecting segment number and polarity in Drosophila. *Nature* **287**, 795–801 (1980).
5. Lewis, E. B. A gene complex controlling segmentation in Drosophila. in *Genes, Development and Cancer* 205–217 (Springer US, 1978).
6. Harding, K., Hoey, T., Warrior, R. & Levine, M. Autoregulatory and gap gene response elements of the even-skipped promoter of Drosophila. *EMBO J.* **8**, 1205–1212 (1989).
7. Goto, T., Macdonald, P. & Maniatis, T. Early and late periodic patterns of even skipped expression are controlled by distinct regulatory elements that respond to different spatial cues. *Cell* **57**, 413–422 (1989).
8. Small, S., Kraut, R., Hoey, T., Warrior, R. & Levine, M. Transcriptional regulation of a pair-rule stripe in Drosophila. *Genes Dev.* **5**, 827–839 (1991).
9. Niessing, D. *et al*. A cascade of transcriptional control leading to axis determination in Drosophila. *J. Cell. Physiol.* **173**, 162–167 (1997).
10. Britten, R. J. & Davidson, E. H. Gene regulation for higher cells: a theory. *Science* **165**, 349–357 (1969).
11. Wray, G. A. The evolutionary significance of cis-regulatory mutations. *Nat. Rev. Genet.* **8**, 206–216 (2007).
12. Shapiro, M. D. *et al*. Genetic and developmental basis of evolutionary pelvic reduction in threespine sticklebacks. *Nature* **428**, 717–723 (2004).
13. Chan, Y. F. *et al*. Adaptive evolution of pelvic reduction in sticklebacks by recurrent deletion of a Pitx1 enhancer. *Science* **327**, 302–305 (2010).
14. Jiang, Q. *et al*. Alzheimer's Disease Variants with the Genome-Wide Significance are Significantly Enriched in Immune Pathways and Active in Immune Cells. *Mol. Neurobiol.* **54**, 594–600 (2017).
15. Farh, K. K.-H. *et al*. Genetic and epigenetic fine mapping of causal autoimmune disease variants. *Nature* **518**, 337–343 (2015).
16. Scott, L. J. *et al*. The genetic regulatory signature of type 2 diabetes in human skeletal muscle. *Nat. Commun.* **7**, 11764 (2016).
17. Maas, S. A. & Fallon, J. F. Single base pair change in the long-range Sonic hedgehog limb-specific enhancer is a genetic basis for preaxial polydactyly. *Dev. Dyn.* **232**, 345–348 (2005).
18. Fleischmann, R. D. *et al*. Whole-genome random sequencing and assembly of Haemophilus influenzae Rd. *Science* **269**, 496–512 (1995).
19. Sayers, E. W. *et al*. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.* **49**, D10–D17 (2021).
20. Adams, M. D. *et al*. The genome sequence of Drosophila melanogaster. *Science* **287**, 2185–2195 (2000).

21. Berman, B. P. *et al*. Exploiting transcription factor binding site clustering to identify cis-regulatory modules involved in pattern formation in the Drosophila genome. *Proc. Natl. Acad. Sci. U. S. A.* **99**, 757–762 (2002).
22. Hallikas, O. *et al*. Genome-wide prediction of mammalian enhancers based on analysis of transcription-factor binding affinity. *Cell* **124**, 47–59 (2006).
23. Shlyueva, D., Stampfel, G. & Stark, A. Transcriptional enhancers: from properties to genome-wide predictions. *Nat. Rev. Genet.* **15**, 272–286 (2014).
24. Hare, E. E., Peterson, B. K., Iyer, V. N., Meier, R. & Eisen, M. B. Sepsid even-skipped enhancers are functionally conserved in Drosophila despite lack of sequence conservation. *PLoS Genet.* **4**, e1000106 (2008).
25. Moses, A. M. *et al*. Large-scale turnover of functional transcription factor binding sites in Drosophila. *PLoS Comput. Biol.* **2**, e130 (2006).
26. Paris, M. *et al*. Extensive divergence of transcription factor binding in Drosophila embryos with highly conserved gene expression. *PLoS Genet.* **9**, e1003748 (2013).
27. Levo, M. *et al*. Unraveling determinants of transcription factor binding outside the core binding site. *Genome Res.* **25**, 1018–1029 (2015).
28. Farley, E. K. *et al*. Suboptimization of developmental enhancers. *Science* **350**, 325–328 (2015).
29. Gordân, R. *et al*. Genomic regions flanking E-box binding sites influence DNA binding specificity of bHLH transcription factors through DNA shape. *Cell Rep.* **3**, 1093–1104 (2013).
30. Buenrostro, J. D., Giresi, P. G., Zaba, L. C., Chang, H. Y. & Greenleaf, W. J. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nat. Methods* **10**, 1213–1218 (2013).
31. Blythe, S. A. & Wieschaus, E. F. Establishment and maintenance of heritable chromatin structure during early Drosophila embryogenesis. *Elife* **5**, (2016).
32. Haines, J. E. & Eisen, M. B. Patterns of chromatin accessibility along the anterior-posterior axis in the early Drosophila embryo. *PLoS Genet.* **14**, e1007367 (2018).
33. Bunina, D. *et al*. Genomic rewiring of SOX2 chromatin interaction network during differentiation of ESCs to postmitotic neurons. *Cell Syst.* **10**, 480-494.e8 (2020).
34. Hannon, C. E., Blythe, S. A. & Wieschaus, E. F. Concentration dependent chromatin states induced by the bicoid morphogen gradient. *Elife* **6**, (2017).
35. Xin, Y. *et al*. Enhancer evolutionary co-option through shared chromatin accessibility input. *Proc. Natl. Acad. Sci. U. S. A.* **117**, 20636–20644 (2020).
36. Siggers, T., Duyzend, M. H., Reddy, J., Khan, S. & Bulyk, M. L. Non-DNA-binding cofactors enhance DNA-binding specificity of a transcriptional regulatory complex. *Mol. Syst. Biol.* **7**, 555 (2011).
37. Spitz, F. & Furlong, E. E. M. Transcription factors: from enhancer binding to developmental control. *Nat. Rev. Genet.* **13**, 613–626 (2012).
38. Calo, E. & Wysocka, J. Modification of enhancer chromatin: what, how, and why? *Mol. Cell* **49**, 825–837 (2013).
39. Boija, A. & Mannervik, M. Initiation of diverse epigenetic states during nuclear programming of theDrosophilabody plan. *Proc. Natl. Acad. Sci. U. S. A.* **113**, 8735–8740 (2016).

40. Schulz, K. N. *et al*. Zelda is differentially required for chromatin accessibility, transcription factor binding, and gene expression in the early Drosophila embryo. *Genome Res*. **25**, 1715–1726 (2015).
41. Li, X.-Y., Harrison, M. M., Villalta, J. E., Kaplan, T. & Eisen, M. B. Establishment of regions of genomic activity during the Drosophila maternal to zygotic transition. *Elife* **3**, (2014).
42. Rickels, R. *et al*. Histone H3K4 monomethylation catalyzed by Trr and mammalian COMPASS-like proteins at enhancers is dispensable for development and viability. *Nat. Genet*. **49**, 1647–1653 (2017).
43. Kellum, R. & Schedl, P. A group of scs elements function as domain boundaries in an enhancer-blocking assay. *Mol. Cell. Biol*. **12**, 2424–2431 (1992).
44. Kellum, R. & Schedl, P. A position-effect assay for boundaries of higher order chromosomal domains. *Cell* **64**, 941–950 (1991).
45. Holdridge, C. & Dorsett, D. Repression of hsp70 heat shock gene transcription by the suppressor of hairy-wing protein of Drosophila melanogaster. *Mol. Cell. Biol*. **11**, 1894–1900 (1991).
46. Schwartz, Y. B. *et al*. Nature and function of insulator protein binding sites in the Drosophila genome. *Genome Res*. **22**, 2188–2198 (2012).
47. Heger, P., George, R. & Wiehe, T. Successive gain of insulator proteins in arthropod evolution. *Evolution* **67**, 2945–2956 (2013).
48. Pugacheva, E. M. *et al*. CTCF mediates chromatin looping via N-terminal domain-dependent cohesin retention. *Proc. Natl. Acad. Sci. U. S. A*. **117**, 2020–2031 (2020).
49. Saldaña-Meyer, R. *et al*. RNA interactions are essential for CTCF-mediated genome organization. *Mol. Cell* **76**, 412-422.e5 (2019).
50. Stadler, M. R., Haines, J. E. & Eisen, M. B. Convergence of topological domain boundaries, insulators, and polytene interbands revealed by high-resolution mapping of chromatin contacts in the early Drosophila melanogaster embryo. *Elife* **6**, (2017).
51. Bonchuk, A. *et al*. Functional role of dimerization and CP190 interacting domains of CTCF protein in Drosophila melanogaster. *BMC Biol*. **13**, 63 (2015).
52. Van Bortle, K. *et al*. Drosophila CTCF tandemly aligns with other insulator proteins at the borders of H3K27me3 domains. *Genome Res*. **22**, 2176–2187 (2012).
53. Bartkuhn, M. *et al*. Active promoters and insulators are marked by the centrosomal protein 190. *EMBO J*. **28**, 877–888 (2009).
54. Gambetta, M. C. & Furlong, E. E. M. The insulator protein CTCF is required for correct Hox gene expression, but not for embryonic development in Drosophila. *Genetics* **210**, 129–136 (2018).
55. Kaushal, A. *et al*. CTCF loss has limited effects on global genome architecture in Drosophila despite critical regulatory functions. *Nat. Commun*. **12**, 1011 (2021).
56. Arzate-Mejía, R. G., Josué Cerecedo-Castillo, A., Guerrero, G., Furlan-Magaril, M. & Recillas-Targa, F. In situ dissection of domain boundaries affect genome topology and gene transcription in Drosophila. *Nat. Commun*. **11**, 894 (2020).
57. Savitsky, M., Kim, M., Kravchuk, O. & Schwartz, Y. B. Distinct roles of chromatin insulator proteins in control of the Drosophila bithorax complex. *Genetics* **202**, 601–617 (2016).

58. Mihaly, J., Hogga, I., Gausz, J., Gyurkovics, H. & Karch, F. In situ dissection of the Fab-7 region of the bithorax complex into a chromatin domain boundary and a Polycomb-response element. *Development* **124**, 1809–1820 (1997).
59. Page, A. R. *et al*. Spotted-dick, a zinc-finger protein of Drosophila required for expression of Orc4 and S phase. *EMBO J*. **24**, 4304–4315 (2005).
60. Cuartero, S., Fresán, U., Reina, O., Planet, E. & Espinàs, M. L. Ibf1 and Ibf2 are novel CP190-interacting proteins required for insulator function. *EMBO J*. **33**, 637–647 (2014).
61. Özdemir, I. & Gambetta, M. C. The Role of Insulation in Patterning Gene Expression. *Genes* **10**, (2019).
62. De Renzis, S., Elemento, O., Tavazoie, S. & Wieschaus, E. F. Unmasking activation of the zygotic genome using chromosomal deletions in the Drosophila embryo. *PLoS Biol*. **5**, e117 (2007).
63. Harrison, M. M. & Eisen, M. B. Transcriptional activation of the zygotic genome in Drosophila. *Curr. Top. Dev. Biol*. **113**, 85–112 (2015).
64. Jukam, D., Shariati, S. A. M. & Skotheim, J. M. Zygotic genome activation in vertebrates. *Dev. Cell* **42**, 316–332 (2017).
65. Harrison, M. M., Li, X.-Y., Kaplan, T., Botchan, M. R. & Eisen, M. B. Zelda binding in the early Drosophila melanogaster embryo marks regions subsequently activated at the maternal-to-zygotic transition. *PLoS Genet*. **7**, e1002266 (2011).
66. Tang, F. *et al*. mRNA-Seq whole-transcriptome analysis of a single cell. *Nat. Methods* **6**, 377–382 (2009).
67. Svensson, V., Vento-Tormo, R. & Teichmann, S. A. Exponential scaling of single-cell RNA-seq in the past decade. *Nat. Protoc*. **13**, 599–604 (2018).
68. Cao, J. *et al*. Comprehensive single-cell transcriptional profiling of a multicellular organism. *Science* **357**, 661–667 (2017).
69. Rosenberg, A. B. *et al*. Single-cell profiling of the developing mouse brain and spinal cord with split-pool barcoding. *Science* **360**, 176–182 (2018).
70. Chari, T. *et al*. Whole Animal Multiplexed Single-Cell RNA-Seq Reveals Plasticity of Clytia Medusa Cell Types. *bioRxiv* 2021.01.22.427844 (2021) doi:10.1101/2021.01.22.427844.
71. Karaiskos, N. *et al*. The Drosophila embryo at single-cell transcriptome resolution. *Science* **358**, 194–199 (2017).
72. Ing-Simmons, E. *et al*. Independence of chromatin conformation and gene regulation during Drosophila dorsoventral patterning. *Nat. Genet*. **53**, 487–499 (2021).
73. Albright, A. R. & Eisen, M. Embryo-to-embryo variability in RNAi knockdown efficiency of dKDM5/lid in Drosophila melanogaster. *bioRxiv* (2021) doi:10.1101/2021.04.29.442033.
74. Luschnig, S. *et al*. An F1 genetic screen for maternal-effect mutations affecting embryonic pattern formation in Drosophila melanogaster. *Genetics* **167**, 325–342 (2004).
75. Staller, M. V. *et al*. Depleting Gene Activities in Early Drosophila Embryos with the "Maternal-Gal4–shRNA" System. *Genetics* **193**, 51–61 (2013).
76. Perrimon, N. Creating mosaics in Drosophila. *Int. J. Dev. Biol*. **42**, 243–247 (1998).
77. Chou, T. B. & Perrimon, N. The autosomal FLP-DFS technique for generating germline mosaics in Drosophila melanogaster. *Genetics* **144**, 1673–1679 (1996).
78. Zirin, J. *et al*. Large-scale transgenic Drosophila resource collections for loss- and gain-of-function studies. *Genetics* **214**, 755–767 (2020).

79. Liang, H.-L. *et al*. The zinc-finger protein Zelda is a key activator of the early zygotic genome in Drosophila. *Nature* **456**, 400–403 (2008).
80. Moshe, A. & Kaplan, T. Genome-wide search for Zelda-like chromatin signatures identifies GAF as a pioneer factor in early fly development. *Epigenetics Chromatin* **10**, (2017).
81. Gaskill, M. M., Gibson, T. J., Larson, E. D. & Harrison, M. M. GAF is essential for zygotic genome activation and chromatin accessibility in the early Drosophila embryo. *Elife* **10**, (2021).
82. Quan, A. S. & Eisen, M. B. The ecology of the Drosophila-yeast mutualism in wineries. *PLoS One* **13**, e0196440 (2018).
83. Luengo Hendriks, C. L. *et al*. Three-dimensional morphology and gene expression in the Drosophila blastoderm at cellular resolution I: data acquisition pipeline. *Genome Biol*. **7**, R123 (2006).
84. Bownes, M. A photographic study of development in the living embryo of Drosophila melanogaster. *J. Embryol. Exp. Morphol*. **33**, 789–801 (1975).
85. Lott, S. E. *et al*. Noncanonical compensation of zygotic X transcription in early Drosophila melanogaster development revealed through single-embryo RNA-seq. *PLoS Biol*. **9**, e1000590 (2011).
86. Wickham, H. *Ggplot2*. (Springer International Publishing, 2016).
87. Zhang, S., Xu, L., Lee, J. & Xu, T. Drosophila atrophin homolog functions as a transcriptional corepressor in multiple developmental processes. *Cell* **108**, 45–56 (2002).
88. Akimaru, H. *et al*. Drosophila CBP is a co-activator of cubitus interruptus in hedgehog signalling. *Nature* **386**, 735–738 (1997).
89. Kronja, I. *et al*. Widespread changes in the posttranscriptional landscape at the Drosophila oocyte-to-embryo transition. *Cell Rep*. **7**, 1495–1508 (2014).
90. Bonn, S. *et al*. Cell type-specific chromatin immunoprecipitation from multicellular complex samples using BiTS-ChIP. *Nat. Protoc*. **7**, 978–994 (2012).
91. Cao, J. *et al*. The single-cell transcriptional landscape of mammalian organogenesis. *Nature* **566**, 496–502 (2019).
92. Gayoso, A. *et al*. Scvi-tools: A library for deep probabilistic analysis of single-cell omics data. *bioRxiv* (2021) doi:10.1101/2021.04.28.441833.
93. Wolf, F. A., Angerer, P. & Theis, F. J. SCANPY: large-scale single-cell gene expression data analysis. *Genome Biol*. **19**, (2018).
94. Melsted, P. *et al*. Modular and efficient pre-processing of single-cell RNA-seq. *bioRxiv* (2019) doi:10.1101/673285.
95. Traag, V. A., Waltman, L. & van Eck, N. J. From Louvain to Leiden: guaranteeing well-connected communities. *Sci. Rep*. **9**, 5233 (2019).
96. McInnes, L., Healy, J. & Melville, J. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. *arXiv [stat.ML]* (2018).
97. Wilcoxon, F. Individual comparisons by ranking methods. *Biom. bull*. **1**, 80 (1945).
98. Hammonds, A. S. *et al*. Spatial expression of transcription factors in Drosophila embryonic organ development. *Genome Biol*. **14**, R140 (2013).
99. Tomancak, P. *et al*. Systematic determination of patterns of gene expression during Drosophila embryogenesis. *Genome Biol*. **3**, RESEARCH0088 (2002).
100. Tomancak, P. *et al*. Global analysis of patterns of gene expression during Drosophila embryogenesis. *Genome Biol*. **8**, R145 (2007).

101. Lex, A., Gehlenborg, N., Strobelt, H., Vuillemot, R. & Pfister, H. UpSet: Visualization of intersecting sets. *IEEE Trans. Vis. Comput. Graph.* **20**, 1983–1992 (2014).
102. Conway, J. R., Lex, A. & Gehlenborg, N. UpSetR: an R package for the visualization of intersecting sets and their properties. *Bioinformatics* **33**, 2938–2940 (2017).
103. Macosko, E. Z. *et al*. Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell* **161**, 1202–1214 (2015).
104. Osorio, D. & Cai, J. J. Systematic determination of the mitochondrial proportion in human and mice tissues for single-cell RNA-sequencing data quality control. *Bioinformatics* **37**, 963–967 (2021).
105. Lopez, R., Regier, J., Cole, M. B., Jordan, M. I. & Yosef, N. Deep generative modeling for single-cell transcriptomics. *Nat. Methods* **15**, 1053–1058 (2018).
106. Skinner, S. O. *et al*. Single-cell analysis of transcription kinetics across the cell cycle. *Elife* **5**, e12175 (2016).
107. Hsiao, C. J. *et al*. Characterizing and inferring quantitative cell cycle phase in single-cell RNA-seq data analysis. *Genome Res.* **30**, 611–621 (2020).
108. McDavid, A., Finak, G. & Gottardo, R. The contribution of cell cycle to heterogeneity in single-cell RNA-seq data. *Nat. Biotechnol.* **34**, 591–593 (2016).
109. Sung, H.-W., Spangenberg, S., Vogt, N. & Großhans, J. Number of nuclear divisions in the Drosophila blastoderm controlled by onset of zygotic transcription. *Curr. Biol.* **23**, 133–138 (2013).
110. Giet, R. & Glover, D. M. Drosophila aurora B kinase is required for histone H3 phosphorylation and condensin recruitment during chromosome condensation and to organize the central spindle during cytokinesis. *J. Cell Biol.* **152**, 669–682 (2001).
111. Wagner, D. E. *et al*. Single-cell mapping of gene expression landscapes and lineage in the zebrafish embryo. *Science* **360**, 981–987 (2018).
112. Briggs, J. A. *et al*. The dynamics of gene expression in vertebrate embryogenesis at single-cell resolution. *Science* **360**, (2018).
113. Farrell, J. A. *et al*. Single-cell reconstruction of developmental trajectories during zebrafish embryogenesis. *Science* **360**, (2018).
114. Hennig, B. P. *et al*. Large-scale low-cost NGS library preparation using a robust Tn5 purification and tagmentation protocol. *G3 (Bethesda)* **8**, 79–89 (2018).
115. Kishi, J. Y., Schaus, T. E., Gopalkrishnan, N., Xuan, F. & Yin, P. Programmable autonomous synthesis of single-stranded DNA. *Nat. Chem.* **10**, 155–164 (2018).
116. Andrews, S. FastQC A Quality Control tool for High Throughput Sequence Data.
117. Jalili, V. *et al*. The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2020 update. *Nucleic Acids Res.* **48**, W395–W402 (2020).
118. Kivioja, T. *et al*. Counting absolute numbers of molecules using unique molecular identifiers. *Nat. Methods* **9**, 72–74 (2011).
119. Islam, S. *et al*. Quantitative single-cell RNA-seq with unique molecular identifiers. *Nat. Methods* **11**, 163–166 (2014).
120. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410 (1990).
121. Telser, A. Molecular biology of the cell, 4th edition. *Shock* **18**, 289 (2002).
122. Zhao, S., Zhang, Y., Gamini, R., Zhang, B. & von Schack, D. Evaluation of two main RNA-seq approaches for gene quantification in clinical RNA sequencing: polyA+ selection versus rRNA depletion. *Sci. Rep.* **8**, (2018).

123. Stapleton, M. *et al*. A Drosophila full-length cDNA resource. *Genome Biol*. **3**, RESEARCH0080 (2002).
124. Aird, D. *et al*. Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries. *Genome Biol*. **12**, R18 (2011).
125. Wimmer, I. *et al*. Systematic evaluation of RNA quality, microarray data reliability and pathway analysis in fresh, fresh frozen and formalin-fixed paraffin-embedded tissue samples. *Sci*. *Rep*. **8**, (2018).
126. Russell, J. N., Clements, J. E. & Gama, L. Quantitation of gene expression in formaldehyde-fixed and fluorescence-activated sorted cells. *PLoS One* **8**, e73849 (2013).
127. Ying, S. Y., Lui, H. M., Lin, S. L. & Chuong, C. M. Generation of full-length cDNA library from single human prostate cancer cells. *Biotechniques* **27**, 410–2, 414 (1999).
128. Wang, X., Yu, L. & Wu, A. R. The effect of methanol fixation on single-cell RNA sequencing data. *BMC Genomics* **22**, 420 (2021).