UNIVERSITY OF CALIFORNIA,
IRVINE

Functional genomic analyses of development in mouse and regeneration in Hydra

DISSERTATION

submitted in partial satisfaction of the requirements
for the degree of

DOCTOR OF PHILOSOPHY

in Biological Sciences

by

Rabi Murad

Dissertation Committee:
Associate Professor Ali Mortazavi, Chair
Professor Robert E. Steele
Professor Ken W. Cho
Professor David M. Gardiner

2018

TABLE OF CONTENTS

LIST OF FIGURES

ACKNOWLEDGMENTS

I would like to thank my graduate advisor Associate Professor Ali Mortazavi for mentoring and supporting me during my graduate studies. I am grateful for the numerous opportunities he provided me that helped my growth as a scientist. His intellect and dedication to science has been a guiding force for me.

I am very grateful to my committee members Professor Robert E. Steele, Professor David M. Gardiner, and Professor Ken W. Cho for their guidance and invaluable advice over the years.

I would like to thank my parents Mohammad Murad and Suraya Aslami, my wife Zeba Ali and my daughters Sophia, Saman, and Sadaf for their love and support. My thesis dissertation would not have been possible without their love and support.

I would like to thank Dr. Debra Mauzy-Melitz for funding support and mentorship. I would also like to thank the staff at Center for Complex Biological Systems, UC Irvine, for their support.

I would like to thank the ENCODE consortium.

I would like to thank all members of Mortazavi Lab for being supportive colleagues and wish all of them best of luck in their professional and personal lives.

CURRICULUM VITAE

Rabi Murad

EDUCATION

University of California, San Diego 2008-2011
Bachelor of Science (B.S.) in Bioengineering (Biotechnology)

University of California, Irvine 2011-2018
Doctor of Philosophy (Ph.D.) in Biological Sciences

RESEARCH AND TRAINEE FELLOWSHIP

1. Predoctoral Training Grant from the Eunice Kennedy Shriver National Institute of Child Health and Human Development (NIH/NICHD) "Training Program in the Systems Biology of Development". Grant Number: 1T32HD060555-05. 2013-2014.

2. Graduate Assistance in Areas of National Need (GAANN) fellowship. Department of Developmental and Cell Biology. University of California, Irvine. 2014-2017.

PUBLICATIONS

1. EV Daniels, **R Murad**, A Mortazavi, RD Reed: Extensive transcriptional response associated with seasonal plasticity of butterfly wing patterns. ***Molecular Ecology*** 2014, 23: 6123-6134.

2. MB Gerstein, J Rozowsky, KK Yan, D Wang, C Cheng, JB Brown, CA Davis, L Hillier, C Sisu, JJ Li, B Pei, AO Harmanci, MO Duff, S Djebali, RP Alexander, BH Alver, R Auerbach, K Bell, PJ Bickel, ME Boeck, NP Boley, BW Booth, L Cherbas, P Cherbas, C Di, A Dobin, J Drenkow, B Ewing, G Fang, M Fastuca, EA Feingold, A Frankish, G Gao, PJ Good, R Guigo, A Hammonds, J Harrow, RA Hoskins, C Howald, L Hu, H Huang, TJ Hubbard, C Huynh, S Jha, D Kasper, M Kato, TC Kaufman, RR Kitchen, E Ladewig, J Lagarde, E Lai, J Leng, Z Lu, M MacCoss, G May, R McWhirter, G Merrihew, DM Miller, A Mortazavi, **R Murad**, *et al*. Comparative analysis of the transcriptome across distant species. ***Nature*** 2014, 512: 445-448.

3. MV Plikus, CF Guerrero-Juarez, M Ito, YR Li, PH Dedhia, Y Zheng, M Shao, DL Gay, R Ramos, TC Hsi, JW Oh, X Wang, A Ramirez, SE Konopelski, A Elzein, A Wang, RJ Supapannachart, HL Lee, CH Lim, A Nace, A Guo, E Treffeisen, T Andl, RN Ramirez, **R Murad**, *et al*. Regeneration of fat cells from myofibroblasts during wound healing. ***Science*** 2017, 355(6326): 748-752.

4. CF Guerrero-Juarez, AA Astrowski, **R Murad**, *et al*. Wound regeneration deficit in rats correlates with low morphogenetic potential and distinct transcriptomic profile of epidermis. ***Journal of Investigative Dermatology*** 2018, 138(6): 1409-1419.

5. L Serra, DZ Chang, M Macchietto, K Williams, **R Murad**, D Lu, AR Dillman, A Mortazavi: Adapting the Smart-seq2 Protocol for Robust Single Worm RNA-seq. ***Bio-protocol*** 2018, 8(4).

6. X Wang, R Ramos, JW Oh, TK Nguyen, HY Liang, V Scarfone, Y Liu, N Taguchi, KN Paolilli, X Wang, G Wang, CF Guerrero-Juarez, S Jiang, R Davis, EN Greenberg, R Ruiz-Vega, S Jahid, P Vasudeva, **R Murad**, *et al*. Signaling by senescent cells hyper-activates the skin stem cell niche. ***Cell*** (under review).

7. **R Murad**, *et al*. Comparative dynamics of microRNA expression during mouse and human prenatal development. (in preparation).

POSTERS

1. Plant and Animal Genomes XXIII (PAG), San Diego, CA, USA, January 2015. "A modular framework for teaching sequencing based functional genomics to high school students" **Rabi Murad**, Marissa Macchietto, Ali Mortazavi, Debra Mauzy-Melitz.

2. 20th Annual International Conference on Intelligent Systems for Molecular Biology (ISMB 2012), Long Beach, CA, USA, July 2012. "microRNA expression levels in ENCODE cell lines" **Rabi Murad**, Weihua Zeng, Ricardo Ramirez, Eddie Park, Ali Mortazavi.

3. Annual retreat, UC Irvine's Center for Complex Biological Systems, Santa Monica, USA, March-April 2012, Poster presentation: "Quantifying the effects of microRNAs on Gene Expression" **Rabi Murad**, Weihua Zeng, Ricardo Ramirez, Eddie Park, Ali Mortazavi.

TALKS & TUTORIALS

20th Annual International Conference on Intelligent Systems for Molecular Biology (ISMB 2012), Long Beach, CA, USA, July 2012. Tutorial: "Analyzing RNA-seq data" **Rabi Murad**, Ali Mortazavi.

TEACHING EXPERIENCE

1. Bio 97: Genetics: Teaching assistant Fall 2012. University of California, Irvine. Supervisors: Dr. Rahul Warrior, Dr. Olivier Cinquin, and Dr. Lee Bardwell

2. Bio D132: Introduction to Personalized Medicine: Teaching assistant Winter 2013. University of California, Irvine. Supervisor: Dr. Ali Mortazavi

3. A Short Course in Systems Biology, Morphogenesis and Spatial Dynamics: Teaching assistant during 2013, 2014, 2015, and 2016. Center for Complex Biological Systems, University of California, Irvine. Supervisor: Dr. Felix Grün

4. California State Summer School for Mathematics and Science (COSMOS): Genes, Genomes, and Biocontrol: Teaching assistant during 2013, 2014, 2015, and 2017. University of California, Irvine. Supervisors: Dr. Ali Mortazavi and Dr. Debra Mauzy-Melitz

5. Bio 93: DNA to Organisms: Teaching assistant Fall 2014, and Fall 2015. University of California, Irvine. Supervisors: Dr. Adrienne Williams and Dr. Justin Shaffer

ABSTRACT OF THE DISSERTATION

Functional genomic analyses of development in mouse and regeneration in Hydra

by

Rabi Murad

Doctor of Philosophy in Biological Sciences

University of California, Irvine, 2018

Associate Professor Ali Mortazavi, Chair

Gene expression at the transcriptional level is controlled by DNA sequences called *cis*-regulatory modules (CRM) and at the post-transcriptional level by microRNAs (miRNAs). CRMs have been studied almost exclusively in bilaterian organisms and little is known about them in non-bilaterian metazoans. Understanding the architecture of CRMs in cnidarians, a sister phylum to bilaterians, can potentially shed light on the evolution of gene regulation. Head regeneration is one of the most widely studied developmental processes in cnidarians. Using a comparison of the transcriptomes of regenerating heads and developing buds, I have determined sets of genes that are specific and common between head regeneration and budding. To understand the genomic sequences controlling these developmental programs, I have mapped the open-chromatin landscape of Hydra in different body parts and during head regeneration to identify candidate promoters and enhancers. My results are the first atlas of CRMs in Hydra, including a substantial fraction that is dynamic during head regeneration.

Mammalian embryonic development has been used as a model system to study the role of miRNAs in previous studies, but a complete atlas of miRNA expression during development is

missing. To understand the role of microRNAs during mouse development, I analyzed a time course of development representing multiple tissues and organs in mouse embryo. We find distinct tissue and developmental stage-specific miRNA expression profiles dominated by a small number of miRNAs. Analysis of conserved miRNAs reveals clustering of expression patterns by tissue types rather than species. We used matching RNA-seq and histone modification ChIP-seq datasets to improve the annotation of miRNA primary transcripts. We show that the expression levels of majority of primary miRNA transcripts predict the expression of their corresponding mature miRNAs. Our data provide the most comprehensive miRNA resource for mouse as well as a comprehensive list of mouse miRNAs that can be reliably measured by RNA-seq of their primary transcripts.

Taken together, the elucidation of cis-regulatory landscape in the cnidarian Hydra and miRNA expression during mouse embryonic development will help the scientific community to understand better the role of enhancers in metazoan evolution and miRNA regulation in mammalian embryonic development.

Chapter 1

Review of enhancer evolution in animals and microRNA expression in bilaterians

1.1 Role of enhancers in metazoan evolution

In 1744 Abraham Trembley carried out probably the first experiment in developmental biology when he bisected a Hydra polyp into two sections and over the next several days observed the cut pieces regenerate missing body parts (Trembley, 1744). Since Trembley's pioneering experiment, Hydra has been used as a versatile model organism for various areas of biological research such as regeneration, aging, stem cell biology, and metazoan evolution (Galliot, 2012). Hydra belongs to the phylum *Cnidaria* that consists of ~10,000 species divided into two major groups: Anthozoa (sea anemones, corals, and sea pens) and Medusozoa (sea wasps, jellyfish, and Hydra). Cnidarians consists of two germ layers (endoderm and ectoderm) and a single body axis (some anthozoans have a second body axis), the anterior-posterior axis also known as the oral-aboral axis.

Cnidarians are of great interest in terms of evolutionary developmental biology since cnidarians and bilaterians diverged about 600 millions ago (Technau & Steele, 2012). Thus, cnidarians form a sister phylum to bilaterians and can potentially provide opportunities for elucidating key aspects of metazoan evolution such as formation of mesoderm, bilaterian body plan, and the nervous system. The availability of genome sequences for metazoans separated by long evolutionary distances provides an opportunity to compare the gene contents, genomic organization, and regulation among these metazoans and study the contribution of each towards metazoan evolution. One explanation considered for greater complexity of bilaterian morphologies and functions compared to cnidarians was a more complex transcriptome in the bilaterians. This idea was put to rest with the sequencing of the genomes of the anthozoan *Nematostella vectensis* (Putnam et al., 2007) and of Hydra (Chapman et al., 2010). Surprisingly the gene contents of *Hydra vulgaris* and *Nematostella vectensis* are similar to those of the bilaterians (Chapman et al., 2010; Putnam et al., 2007). This finding led to speculation that the

difference in the body plans of cnidarians and bilaterians is due to complexity of gene regulation (Schwaiger, 2014) based on findings that body plan evolution is often a consequence of changes in gene regulation (Carroll, 2008).

Development of a complex multicellular organism from a single-cell zygote to adult form is achieved by precise spatio-temporal expression of specific subsets of genes orchestrated by gene regulatory sequences called enhancers. While enhancer function and evolution have largely been studied in the context of mammalian and a few other bilaterian model systems (Consortium et al., 2012; Neph et al., 2012; Roy et al., 2010; Visel et al., 2013; Visel, Rubin, & Pennacchio, 2009), few studies are reported in plants, fungi, and non-bilaterian organisms. Specifically, no attempt has been made to decipher the role of enhancers in the evolution of complex bilaterians from basal organisms in the metazoan kingdom.

The sequencing and comparative analysis of genomes of closely related animals, such as human and chimpanzee, reveals that there is high degree of conservation at the DNA level (Mikkelsen et al., 2005) and yet there are marked differences that distinguish the two species. Similarly, conservation even among distantly related species is high at the protein level. The defining feature distinguishing closely related animals, such as human from apes, is differences in gene expression patterns (King & Wilson, 1975). These findings hint at evolutionary innovation occurring more often in regulatory sequences than in the gene sequences. Indeed comparative studies of selected mammals reveal that enhancers may evolve more rapidly than the gene expression patterns during evolution (Cotney et al., 2013; Xiao et al., 2012). A more recent study shows that enhancer evolution is a common feature of mammalian genomes (Villar et al., 2015).

Although the above studies in mammalian model systems show that enhancers evolve at a more rapid pace than the coding sequences, it remains to be determined if enhancer evolution is an important agent of metazoan evolution over long evolutionary periods. Mapping the gene regulatory elements and their functions in basal metazoans such as cnidarians and comparing them to the bilaterians can answer this question.

Here we review general features of metazoan enhancers, experimental approaches to mapping the enhancer regions genome-wide, evolution of enhancers, and how the evolution of enhancers can contribute to the evolution of metazoan form and function.

Genomic features of metazoan enhancers

Gene expression at the transcriptional level is controlled by DNA sequences called *cis*-regulatory modules (CRM). CRMs are short genomic segments that play specific roles in controlling the expression of their associated genes. CRMs act as binding sites for sequence-specific as well as general transcription factors (TFs) that mediate the expression of genes. CRMs are classified into three groups: promoters, insulators, and enhancers. Promoters are short genomic regions overlapping the transcription start sites (TSSs) of genes and act as binding sites for general transcription factors to mediate the binding of RNA Polymerase II (Maston, Evans, & Green, 2006). Insulators are CRMs that block the activity of enhancers and therefore reduce gene expression (Chung, Whiteley, & Felsenfeld, 1993). Enhancers are *cis*-acting 200-1000 base pair (bp) genomic regions that contain multiple transcription factor binding sites (TFBSs) and control the expression of nearby target genes (Levine, 2010).

The first enhancer was identified in the SV40 animal virus consisting of two 72 bp repeats which are ~200 bp upstream of the T-antigen gene essential for viral replication in

infected cells (Banerji, Rusconi, & Schaffner, 1981). This finding revealed for the first time the complex eukaryotic gene regulatory organization in which the regulatory sequences in the form of enhancers are decoupled from the core promoters (Levine, 2010). They are usually located tens to hundreds of kilobases (Kb) from their associated genes (Levine, 2010) and can be located up to 1 megabase (Mb) from their targets (Lettice et al., 2003). Once the right combination of transcription factors are bound to the TFBSs at an enhancer, the enhancer region is brought within close proximity of the target gene promoter by DNA looping (Bulger & Groudine, 1999).

Approaches to identifying enhancers in metazoan genomes

The first step in a comparative study of enhancers is the genome-wide identification of enhancers in the model organisms of interest. Three major approaches have been used for identifying CRMs with varying degrees of success (Hardison & Taylor, 2012). The first two approaches use bioinformatics methods to scan for short genomic segments that possess some enhancer characteristics, while the third approach uses experimental techniques to determine the genomic enhancer locations genome-wide.

The first approach involves scanning short genomic regions for clusters of transcription factor binding sites (TFBS) provided that the TF motif sequences are known (Hardison & Taylor, 2012). This approach leads to many false positives since clusters of TFBSs are abundant in genomes (Hardison & Taylor, 2012). The second approach relies on comparing the non-coding genomic regions of interest for signs of evolutionary conservation corresponding to enhancers (Hardison & Taylor, 2012). This approach fails to identify lineage-specific enhancers that are not conserved among evolutionarily distant organisms.

The third and the most reliable approach involves experimental methods to identify genomic regions containing features associated with enhancers. Experimental methods for mapping the CRMs are based on detecting combinations of histone modifications that are associated with each CRM type and the "open chromatin" accessibility of CRM regions due to localized depletion of nucleosomes.

Specific combinations of posttranslational modifications mark histone proteins at each CRM subtype. The specific histone modifications ("marks") can be used to identify the different types of CRMs. For example, the histone proteins in promoter regions are marked by high levels of H3K4me3 and H3K4me2, while high levels of H3K4me2 and K3K4me1 (low levels of H3K4me3) mark enhancer regions. H3K27ac mark is associated with active promoters and enhancers. Chromatin immunoprecipitation with an antibody recognizing a specific histone mark followed by sequencing can be used to map the locations of specific CRM types genome-wide (Visel, Blow, et al., 2009). This experimental approach can have high predictive power for CRMs depending on the quality of the antibodies used.

Active enhancers (and all non-repressed CRMs generally) are located in genomic regions devoid of nucleosomes so that the TFBSs can be accessed by transcription factors. Such nucleosome-free regions are called open chromatin regions. Open chromatin regions are prone to enzymatic digestion due to easy accessibility. This property has been used in developing high-throughput assays, such as DNase-seq (Boyle et al., 2011; Hesselberth et al., 2009; Neph et al., 2012) and ATAC-seq (Buenrostro, Giresi, Zaba, Chang, & Greenleaf, 2013), which preferentially digest open-chromatin regions with enzymes DNase and Tn5 transposase respectively. ATAC-seq has gained popularity over DNase-seq due to the ease of the experimental protocol and low cell number requirement (500-50000 cells) (Buenrostro et al.,

2013).

Evolution of metazoan enhancers

Highly conserved enhancers are known to regulate important processes such as embryonic development (Pennacchio et al., 2006), but enhancer evolution has been shown to underlie evolutionary differences among metazoans (Cotney et al., 2013; Villar et al., 2015). Although enhancers evolve at a slower place compared to non-functional DNA because enhancer sequences are subject to purifying evolutionary selection, they evolve faster than the protein coding sequences that they control through both point and indel mutations (Rubinstein & de Souza, 2013). Drastic evolution of enhancers can also occur when new enhancers are born through insertion of transposable elements, *de novo* mutations in non-regulatory sequences, or chromosomal rearrangements, whereas enhancers are lost through deletion of enhancer segment for instance (Rubinstein & de Souza, 2013). A more subtle form of enhancer evolution occurs through point mutations and small indels to fine-tune the affinity of TFBSs within the enhancers (Rubinstein & de Souza, 2013). Availability of genome sequences and catalogues of regulatory sequences for well-studied model organisms has allowed comparative analysis of enhancer evolution. A comparative analysis of TFBSs between mouse and human has revealed that the conservation of individual TFBSs is low (22%) although the overall architecture of regulatory networks is very conserved (> 95%) between the two species (Stergachis et al., 2014).

The availability of genome sequences in cnidarians has enabled us to ask questions about the evolution of gene regulatory landscapes from basal metazoans to the bilaterians.

1.2 The role of microRNAs in mammalian embryonic development

microRNAs (miRNAs) are small ~22 nucleotide (nt) endogenous non-coding RNAs that regulate gene expression by mediating the post-transcriptional degradation of messenger RNA (mRNA) or hindering the translation of proteins by acting as guide RNAs for the protein Argonaute (Ago) (Bartel, 2004). *lin-4*, the first miRNA discovered, was found to regulate developmental timing in the nematode *C. elegans* (Lee, Feinbaum, & Ambros, 1993). The miRNA *let-7* was also found to regulate developmental timing of *C. elegans* (Reinhart BJ1, Slack FJ, Basson M, Pasquinelli AE, Bettinger JC, Rougvie AE & HR, 2000). Subsequently it became apparent that miRNA-mediated regulation of gene expression is not limited to nematodes but is present in diverse metazoans (Pasquinelli et al., 2000).

miRNAs are regulators of diverse cellular processes during development and homeostasis (Vidigal & Ventura, 2015), and their dysregulation is known to underlie many diseases and developmental defects (Bartel, 2018). With increasing interest in miRNA biology and the use of cloning as well as the rise of sequencing technologies, miRNAs have been profiled and studied in diverse species. Consequently, hundreds of miRNAs have been discovered and annotated in many metazoans, plants, and viruses (Kozomara & Griffiths-Jones, 2011). For example, hundreds of human miRNAs have been discovered, most of which are conserved in diverse organisms (Bartel, 2018).

Beyond the discovery of miRNAs in different model organisms, many studies have investigated miRNA expression levels in diverse cells and tissues, both in normal and disease states. Such efforts have led to deep understanding of tissue- and cell-specific miRNA expression patterns. For example, Mineno and colleagues used massively parallel signature sequencing (MPSS) technology to profile miRNAs in mouse embryos during three embryonic stages (e9.5, e10.5, and e11.5) and were able to detect 390 distinct miRNAs (Mineno et al., 2006). Chiang and

colleagues extended this work by sequencing small RNAs from mouse brain, ovary, testes, embryonic stem cells, embryonic stages of complete embryos from three developmental stages, and whole newborns to profile the expression of 398 annotated and 108 novel miRNAs (Chiang et al., 2010). Landgraf and colleagues cloned and sequenced more than 250 small RNA libraries from 26 different organ and cell types from humans and rodents to profile miRNA expression and describe various other miRNA characteristics (Landgraf et al., 2007). More recently, the FANTOM5 project created a miRNA expression atlas using deep-sequencing data from 396 human and 47 mouse RNA samples (De Rie et al., 2017). However, a complete and systematic atlas of miRNA expression in tissues representative of major organ systems and broad number of mammalian embryonic stages is still missing. To fill this knowledge gap we profiled the expression of miRNAs in 16 different tissues during a time-course of 8 embryonic stages in mouse as well as in several human prenatal samples using multiple complementary sequencing and hybridization techniques to provide a more complete mammalian embryonic miRNA expression atlas for the scientific community.

Here we review miRNA biogenesis, targeting, expression profiling techniques, and their myriad roles in embryonic development.

miRNA biogenesis

miRNA biogenesis may proceed through a canonical pathway or non-canonical pathway. Canonical miRNA biogenesis occurs in several steps. Polyadenylated primary miRNA (pri-miRNA) transcripts (>200 nt), which are sometimes referred to as “host genes” and have a characteristic hairpin structure, are transcribed by RNA polymerase II. The pri-miRNA transcript may host a single or multiple miRNA hairpin structures called poylcistronic miRNA transcripts.

The hairpin(s) in the pri-miRNA transcript are cleaved in the nucleus by the enzyme Drosha into pre-miRNAs (~80 nt) that are exported to the cytoplasm and finally processed into 21-24 nt mature miRNAs by the enzyme Dicer (Han et al., 2006). An example of noncanonical miRNA biogenesis is the processing of "mirtrons" which are transcripts of introns that enter the miRNA biogenesis pathway as pre-miRNAs after their ends are cleaved by spliceosome (Okamura, Hagen, Duan, Tyler, & Lai, 2007).

miRNA targeting and target prediction

miRNA target recognition occurs by Watson-Crick pairing of the miRNA seed region (the first 2-7 nucleotides) and the miRNA binding site in the target mRNA which is usually located in the 3' UTR (Bartel, 2009). miRNA binding sites are very abundant in the target mRNA 3' UTRs, with each human mRNA 3' UTR often containing more than 300 binding sites for miRNAs (Friedman, Farh, Burge, & Bartel, 2009). Also, a majority of human mRNAs (~60%) are targeted by miRNAs (Friedman et al., 2009).

Reliable prediction of miRNA targets is important for deciphering the regulatory role of miRNAs. Algorithms have been developed that search for miRNA binding sites in target mRNAs and have led to predicted sets of miRNA targets in several organisms (Agarwal, Bell, Nam, & Bartel, 2015; Bartel, 2009). Further assessment of these predicted miRNA targets is needed to filter out false-positives, and there may be some miRNA targets that are missed. Such assessment can be done computationally or experimentally. A computational method involves assessing the evolutionary conservation of miRNA target sites across multiple species (Bartel, 2009). Experimental miRNA-target identification can be carried out by crosslinking the miRNA effector protein Ago to the miRNA targets followed by a sequencing technique called eCLIP

(Van Nostrand et al., 2016). An alternative technique called CLASH involves the ligation of miRNAs and their target mRNAs followed by sequencing (Helwak, Kudla, Dudnakova, & Tollervey, 2013). A simpler computational approach that can assess miRNA-target predictions can be developed if the expression levels of miRNAs and their predicted targets are available across multiple cells and/or tissues. Observing and quantifying the correlation of the expressions of the miRNAs and their predicted targets can be used to identify the miRNA targets.

miRNA profiling techniques

With the growing evidence of the critical role of miRNAs in homeostasis and disease, multiple experimental techniques have been developed to profile the expression of mature miRNAs, each with their own strengths (Mestdagh et al., 2014). RNA-seq typically refers to profiling expressed transcripts 200 nt or longer including messenger RNAs (mRNAs) and long non-coding RNAs (lncRNAs) (Mortazavi, Williams, McCue, Schaeffer, & Wold, 2008), which here we will refer to as messenger RNA-seq (mRNA-seq), whereas there are also multiple miRNA-specific sequencing protocols such as microRNA-seq (Roberts et al., 2015) and short RNA-seq (Fejes-Toth et al., 2009). There are hybridization-based assays such as microarrays as well as molecule counting such as NanoString, which involves hybridization of color-coded molecular barcodes (Geiss et al., 2008; Wyman et al., 2011). As mature miRNAs are processed from longer pri-miRNAs and the annotated pri-miRNAs are predominantly protein-coding or lncRNA transcripts (Cai, Hagedorn, & Cullen, 2004), in theory mRNA-seq should be able to profile the expression of pri-miRNAs. However, there is a significant number of miRNAs whose host genes (i.e., genes that give rise to their respective pri-miRNAs) have not been annotated. Furthermore, an important question remains whether the expression of pri-miRNAs can reliably

predict the expression of their corresponding mature miRNAs. This would allow the simultaneous expression profiling of the mature miRNA along with mRNAs using mRNA-seq. Previous studies have attempted to answer this question in specific cell types (Zeng et al., 2016). Availability of matching mRNA-seq and microRNA-seq data sets for the same samples provides a unique opportunity to answer the following question: Can the expression of mature miRNAs be reliably predicted from the expression of their pri-miRNA transcripts?

Role of miRNAs in metazoan development

The first two miRNAs discovered, *lin-4* and *let-7*, were found to be important regulators of developmental timing in *C. elegans* (Lee et al., 1993; Reinhart BJ1, Slack FJ, Basson M, Pasquinelli AE, Bettinger JC, Rougvie AE & HR, 2000). Since their initial discovery, many other miRNAs have been shown to be involved in developmental processes. Loss-of-function studies have revealed that miRNA dysregulation leads to developmental defect phenotypes in skeleton, teeth, brain, eyes, neurons, muscle, heart, lungs, kidneys, vasculature, liver, pancreas, intestine, skin, fat, breast, ovaries, testes, placenta, thymus, and each hematopoietic lineage, as well as cellular, physiological, and behavioral defects (Bartel, 2018). Knock-down of a single miRNA usually does not lead to severe developmental defects since most of the miRNAs share seed regions with the members of their families that provide functional redundancy (Bartel, 2018). That being said, there are a few exceptions. In mice, knock-down of miRNA-128-2, a brain-specific miRNA, led to development of fatal epilepsy (Tan et al., 2013) while the knock-down of four members of the miR-291a/294/302abcd family (miR-302a, miR-302b, miR-302c, and miR302d) led to late embryonic lethality (Parchem et al., 2015). This limited number of

examples reveals that miRNAs play an important role in development and much can be learnt about their biological functions through their characterization and functional analysis.

A comprehensive atlas of miRNAs during embryonic development with information about their tissue-specific and temporal expression dynamics, their targets, and their conservation of sequence and function will help form a complete picture of the role of miRNAs in normal development and etiology of developmental defects. Towards this goal we profiled the expression of miRNAs in 16 tissues encompassing all major organ systems during a time-course of 8 embryonic stages (e10.5 – P0) in mouse (Chapter 4).

References

Agarwal, V., Bell, G. W., Nam, J. W., & Bartel, D. P. (2015). Predicting effective microRNA target sites in mammalian mRNAs. *ELife*, *4*(AUGUST2015), 1–38. https://doi.org/10.7554/eLife.05005

Banerji, J., Rusconi, S., & Schaffner, W. (1981). Expression of a β-globin gene is enhanced by remote SV40 DNA sequences. *Cell*, *27*(2 PART 1), 299–308. https://doi.org/10.1016/0092-8674(81)90413-X

Bartel, D. P. (2004). MicroRNAs: Genomics, Biogenesis, Mechanism, and Function. *Cell*, *116*(2), 281–297. https://doi.org/10.1016/S0092-8674(04)00045-5

Bartel, D. P. (2009). MicroRNAs: Target Recognition and Regulatory Functions. *Cell*, *136*(2), 215–233. https://doi.org/10.1016/j.cell.2009.01.002

Bartel, D. P. (2018). Metazoan MicroRNAs. *Cell*, *173*(1), 20–51. https://doi.org/10.1016/j.cell.2018.03.006

Boyle, A. P., Song, L., Lee, B. K., London, D., Keefe, D., Birney, E., … Furey, T. S. (2011). High-resolution genome-wide in vivo footprinting of diverse transcription factors in human cells. *Genome Research*, *21*(3), 456–464. https://doi.org/10.1101/gr.112656.110

Buenrostro, J. D., Giresi, P. G., Zaba, L. C., Chang, H. Y., & Greenleaf, W. J. (2013). Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nature Methods*, *10*(12), 1213–1218. https://doi.org/10.1038/nmeth.2688

Bulger, M., & Groudine, M. (1999). Looping versus linking: Toward a model for long-distance gene activation. *Genes and Development*, *13*(19), 2465–2477. https://doi.org/10.1101/gad.13.19.2465

Cai, X., Hagedorn, C. H., & Cullen, B. R. (2004). Human microRNAs are processed from capped, polyadenylated transcripts that can also function as mRNAs. *RNA (New York, N.Y.)*, *10*(12), 1957–66. https://doi.org/10.1261/rna.7135204

Carroll, S. B. (2008). Evo-Devo and an Expanding Evolutionary Synthesis: A Genetic Theory of Morphological Evolution. *Cell*, *134*(1), 25–36. https://doi.org/10.1016/j.cell.2008.06.030

Chapman, J. A., Kirkness, E. F., Simakov, O., Hampson, S. E., Mitros, T., Weinmaier, T., … Steele, R. E. (2010). The dynamic genome of Hydra. *Nature*, *464*(7288), 592–596. https://doi.org/10.1038/nature08830

Chiang, H. R., Schoenfeld, L. W., Ruby, J. G., Auyeung, V. C., Spies, N., Baek, D., … Bartel, D. P. (2010). Mammalian microRNAs: Experimental evaluation of novel and previously annotated genes. *Genes and Development*, *24*(10), 992–1009. https://doi.org/10.1101/gad.1884710

Chung, J. H., Whiteley, M., & Felsenfeld, G. (1993). A 5′ element of the chicken β-globin domain serves as an insulator in human erythroid cells and protects against position effect in Drosophila. *Cell*, *74*(3), 505–514. https://doi.org/10.1016/0092-8674(93)80052-G

Consortium, E. P., Dunham, I., Kundaje, A., Aldred, S. F., Collins, P. J., Davis, C. a, … Lochovsky, L. (2012). An integrated encyclopedia of DNA elements in the human genome. *Nature*, *489*(7414), 57–74. https://doi.org/10.1038/nature11247

Cotney, J., Leng, J., Yin, J., Reilly, S. K., Demare, L. E., Emera, D., … Noonan, J. P. (2013). XThe evolution of lineage-specific regulatory activities in the human embryonic limb. *Cell*, *154*(1), 185–196. https://doi.org/10.1016/j.cell.2013.05.056

De Rie, D., Abugessaisa, I., Alam, T., Arner, E., Arner, P., Ashoor, H., … De Hoon, M. J. L.

(2017). An integrated expression atlas of miRNAs and their promoters in human and mouse. *Nature Biotechnology*, *35*(9), 872–878. https://doi.org/10.1038/nbt.3947

Fejes-Toth, K., Sotirova, V., Sachidanandam, R., Assaf, G. Hannon, G. J., Kapranov, P., Foissac, S., … Gingeras, T. R. (2009). Post-transcriptional processing generates a diversity of 5'-modified long and short RNAs. *Nature*, *457*(7232), 1028–1032. https://doi.org/10.1038/nature07759

Friedman, R. C., Farh, K. K. H., Burge, C. B., & Bartel, D. P. (2009). Most mammalian mRNAs are conserved targets of microRNAs. *Genome Research*, *19*(1), 92–105. https://doi.org/10.1101/gr.082701.108

Galliot, B. (2012). Hydra , a fruitful model system for 270 years, *423*(July), 411–423. https://doi.org/10.1387/ijdb.120086bg

Geiss, G. K., Bumgarner, R. E., Birditt, B., Dahl, T., Dowidar, N., Dunaway, D. L., … Dimitrov, K. (2008). Direct multiplexed measurement of gene expression with color-coded probe pairs. *Nature Biotechnology*, *26*(3), 317–25. https://doi.org/10.1038/nbt1385

Han, J., Lee, Y., Yeom, K. H., Nam, J. W., Heo, I., Rhee, J. K., … Kim, V. N. (2006). Molecular Basis for the Recognition of Primary microRNAs by the Drosha-DGCR8 Complex. *Cell*, *125*(5), 887–901. https://doi.org/10.1016/j.cell.2006.03.043

Hardison, R. C., & Taylor, J. (2012). Genomic approaches towards finding cis-regulatory modules in animals. *Nature Reviews Genetics*, *13*(7), 469–483. https://doi.org/10.1038/nrg3242

Helwak, A., Kudla, G., Dudnakova, T., & Tollervey, D. (2013). Mapping the human miRNA interactome by CLASH reveals frequent noncanonical binding. *Cell*, *153*(3), 654–665. https://doi.org/10.1016/j.cell.2013.03.043

Hesselberth, J. R., Chen, X., Zhang, Z., Sabo, P. J., Sandstrom, R., Reynolds, A. P., … Stamatoyannopoulos, J. A. (2009). Global mapping of protein-DNA interactions in vivo by digital genomic footprinting. *Nature Methods*, *6*(4), 283–289. https://doi.org/10.1038/nmeth.1313

King, M. C., & Wilson, A. C. (1975). Evolution at two levels in humans and chimpanzees. *Science (New York, N.Y.)*, *188*(4184), 107–116. https://doi.org/10.1126/science.1090005

Kozomara, A., & Griffiths-Jones, S. (2011). MiRBase: Integrating microRNA annotation and deep-sequencing data. *Nucleic Acids Research*, *39*(SUPPL. 1), 152–157. https://doi.org/10.1093/nar/gkq1027

Landgraf, P., Rusu, M., Sheridan, R., Sewer, A., Iovino, N., Aravin, A., … Tuschl, T. (2007). A Mammalian microRNA Expression Atlas Based on Small RNA Library Sequencing. *Cell*, *129*(7), 1401–1414. https://doi.org/10.1016/j.cell.2007.04.040

Lee, R. C., Feinbaum, R. L., & Ambros, V. (1993). the C. elegans\rheterochronic gene lin-4 encodes small RNAs with antisense\rcomplementarity to lin-14. *Cell* , *75: 843–85*, 843–854. https://doi.org/10.1016/0092-8674(93)90529-Y

Lettice, L. A., Heaney, S. J. H., Purdie, L. A., Li, L., de Beer, P., Oostra, B. A., … de Graaff, E. (2003). A long-range Shh enhancer regulates expression in the developing limb and fin and is associated with preaxial polydactyly. *Human Molecular Genetics*, *12*(14), 1725–1735. https://doi.org/10.1093/hmg/ddg180

Levine, M. (2010). Transcriptional enhancers in animal development and evolution. *Current Biology*, *20*(17), R754–R763. https://doi.org/10.1016/j.cub.2010.06.070

Maston, G. A., Evans, S. K., & Green, M. R. (2006). Transcriptional Regulatory Elements in the Human Genome. *Annual Review of Genomics and Human Genetics*, *7*(1), 29–59.

https://doi.org/10.1146/annurev.genom.7.080505.115623
Mestdagh, P., Hartmann, N., Baeriswyl, L., Andreasen, D., Bernard, N., Chen, C., … Vandesompele, J. (2014). Evaluation of quantitative miRNA expression platforms in the microRNA quality control (miRQC) study. *Nature Methods*, *11*(8), 809–815. https://doi.org/10.1038/nmeth.3014
Mikkelsen, T. S., Hillier, L. W., Eichler, E. E., Zody, M. C., Jaffe, D. B., Yang, S. P., … Waterston, R. H. (2005). Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature*, *437*(7055), 69–87. https://doi.org/10.1038/nature04072
Mineno, J., Okamoto, S., Ando, T., Sato, M., Chono, H., Izu, H., … Kato, I. (2006). The expression profile of microRNAs in mouse embryos. *Nucleic Acids Research*, *34*(6), 1765–1771. https://doi.org/10.1093/nar/gkl096
Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L., & Wold, B. (2008). Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nature Methods*, *5*(7), 621–628. https://doi.org/10.1038/nmeth.1226
Neph, S., Vierstra, J., Stergachis, A. B., Reynolds, A. P., Haugen, E., Vernot, B., … Stamatoyannopoulos, J. A. (2012). An expansive human regulatory lexicon encoded in transcription factor footprints. *Nature*, *489*(7414), 83–90. https://doi.org/10.1038/nature11212
Okamura, K., Hagen, J. W., Duan, H., Tyler, D. M., & Lai, E. C. (2007). The Mirtron Pathway Generates microRNA-Class Regulatory RNAs in Drosophila. *Cell*, *130*(1), 89–100. https://doi.org/10.1016/j.cell.2007.06.028
Parchem, R. J., Moore, N., Fish, J. L., Parchem, J. G., Braga, T. T., Shenoy, A., … Blelloch, R. (2015). miR-302 Is Required for Timing of Neural Differentiation, Neural Tube Closure, and Embryonic Viability. *Cell Reports*, *12*(5), 760–773. https://doi.org/10.1016/j.celrep.2015.06.074
Pasquinelli, A. E., Reinhart, B. J., Slack, F., Martindale, M. Q., Kuroda, M. I., Maller, B., … Ruvkun, G. (2000). Conservation of the sequence and temporal expression of let-7\nheterochronic regulatory RNA. *Nature*, *408*(6808), 86–89. https://doi.org/10.1038/35040556
Patro, R., Duggal, G., Love, M. I., Irizarry, R. A., & Kingsford, C. (2017). Salmon provides fast and bias-aware quantification of transcript expression. *Nature Methods*, *14*(4), 417–419. https://doi.org/10.1038/nmeth.4197
Pennacchio, L. A., Ahituv, N., Moses, A. M., Prabhakar, S., Nobrega, M. A., Shoukry, M., … Rubin, E. M. (2006). In vivo enhancer analysis of human conserved non-coding sequences. *Nature*, *444*(7118), 499–502. https://doi.org/10.1038/nature05295
Putnam, N. H., Srivastava, M., Hellsten, U., Dirks, B., Chapman, J., Salamov, A., … Rokhsar, D. S. (2007). Sea Anemone Genome Reveals Ancestral Eumetazoan Gene Repertoire and Genomic Organization. *Science*, *317*(5834), 86–94. https://doi.org/10.1126/science.1139158
Reinhart BJ1, Slack FJ, Basson M, Pasquinelli AE, Bettinger JC, Rougvie AE, H., & HR, R. G. (2000). The 21-nucleotide let- 7 RNA regulates developmental timing in Caenorhabditis. *Nature*, *403*(6772), 901–6.
Roberts, B. S., Hardigan, A. A., Kirby, M. K., Fitz-Gerald, M. B., Wilcox, C. M., Kimberly, R. P., & Myers, R. M. (2015). Blocking of targeted microRNAs from next-generation sequencing libraries. *Nucleic Acids Research*, *43*(21), 1–8. https://doi.org/10.1093/nar/gkv724
Roy, S., Ernst, J., Kharchenko, P. V., Kheradpour, P., Negre, N., Eaton, M. L., … Lowdon, R. F.

(2010). Identification of functional elements and regulatory circuits by Drosophila modENCODE. *Science*, *330*(6012), 1787–1797. https://doi.org/10.1126/science.1198374

Rubinstein, M., & de Souza, F. S. J. (2013). Evolution of transcriptional enhancers and animal diversity. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, *368*(1632), 20130017. https://doi.org/10.1098/rstb.2013.0017

Schwaiger, M. (2014). Evolutionary conservation of the eumetazoam gene regulatory landscape - Supplemental Figures. *Genome Research*, 1–13. https://doi.org/10.1101/gr.162529.113.Freely

Stergachis, A. B., Neph, S., Sandstrom, R., Haugen, E., Reynolds, A. P., Zhang, M., … Stamatoyannopoulos, J. A. (2014). Conservation of trans-acting circuitry during mammalian regulatory evolution. *Nature*, *515*(7527), 365–370. https://doi.org/10.1038/nature13972

Tan, C. L., Plotkin, J. L., Venø, M. T., Von Schimmelmann, M., Feinberg, P., Mann, S., … Schaefer, A. (2013). MicroRNA-128 governs neuronal excitability and motor behavior in mice. *Science*, *342*(6163), 1254–1258. https://doi.org/10.1126/science.1244193

Technau, U., & Steele, R. E. (2012). Evolutionary crossroads in developmental biology: Cnidaria. *Development*, *139*(23), 4491–4491. https://doi.org/10.1242/dev.090472

Van Nostrand, E. L., Pratt, G. A., Shishkin, A. A., Gelboin-Burkhart, C., Fang, M. Y., Sundararaman, B., … Yeo, G. W. (2016). Robust transcriptome-wide discovery of RNA-binding protein binding sites with enhanced CLIP (eCLIP). *Nature Methods*, *13*(6), 508–514. https://doi.org/10.1038/nmeth.3810

Vidigal, J. A., & Ventura, A. (2015). The biological functions of miRNAs: Lessons from in vivo studies. *Trends in Cell Biology*, *25*(3), 137–147. https://doi.org/10.1016/j.tcb.2014.11.004

Villar, D., Berthelot, C., Aldridge, S., Rayner, T. F., Lukk, M., Pignatelli, M., … Odom, D. T. (2015). Enhancer evolution across 20 mammalian species. *Cell*, *160*(3), 554–566. https://doi.org/10.1016/j.cell.2015.01.006

Visel, A., Blow, M. J., Li, Z., Zhang, T., Akiyama, J. A., Holt, A., … Pennacchio, L. A. (2009). ChIP-seq accurately predicts tissue-specific activity of enhancers. *Nature*, *457*(7231), 854–858. https://doi.org/10.1038/nature07730

Visel, A., Rubin, E. M., & Pennacchio, L. A. (2009). Genomic views of distant-acting enhancers. *Nature*. https://doi.org/10.1038/nature08451

Visel, A., Taher, L., Girgis, H., May, D., Golonzhka, O., Hoch, R. V., … Rubenstein, J. L. R. (2013). A high-resolution enhancer atlas of the developing telencephalon. *Cell*, *152*(4), 895–908. https://doi.org/10.1016/j.cell.2012.12.041

Wyman, S. K., Knouf, E. C., Parkin, R. K., Fritz, B. R., Lin, D. W., Dennis, L. M., … Tewari, M. (2011). Post-transcriptional generation of miRNA variants by multiple nucleotidyl transferases contributes to miRNA transcriptome complexity. *Genome Research*, *21*(9), 1450–1461. https://doi.org/10.1101/gr.118059.110

Xiao, S., Xie, D., Cao, X., Yu, P., Xing, X., Chen, C. C., … Zhong, S. (2012). Comparative epigenomic annotation of regulatory DNA. *Cell*, *149*(6), 1381–1392. https://doi.org/10.1016/j.cell.2012.04.029

Zeng, W., Jiang, S., Kong, X., El-ali, N., Ball, A. R., Ma, C. I.-H., … Mortazavi, A. (2016). Single-nucleus RNA-seq of differentiating human myoblasts reveals the extent of fate heterogeneity. *Nucleic Acids Research*, 1–13. https://doi.org/10.1093/nar/gkw739

Chapter 2

Comparative dynamics of the transcriptome during regeneration and budding in Hydra

Statement of contribution

In this study, I performed experiments, analyzed data and interpreted results. Ashley Wong (currently at Medical College of Wisconsin in Milwaukee) contributed technically with tissue manipulation, RNA extraction and RNA-seq of budding stages.

2.1 Abstract

Hydra has long been studied for its remarkable ability to regenerate, which is controlled by a set of cells called the head organizer near the hypostome. Previous studies that focused on the molecular mechanisms of axial patterning and head regeneration in Hydra have revealed the role of the canonical Wnt pathway in the Hydra head organizer. The canonical Wnt pathway is also expressed in the organizer of a developing bud, the asexual mode of reproduction in Hydra. However, it is unclear how shared the developmental programs of head organizer genesis are in budding and regeneration. We therefore performed a global gene expression analysis during Hydra head regeneration and budding time-courses using RNA-seq to answer this question. Time-series analysis of head regeneration and budding revealed a set of 6305 differentially expressed transcripts in thirteen clusters with distinct expression profiles during the 48-hour head regeneration and 72-hour budding time-courses. Three of these clusters are shared between budding and head regeneration while the rest are specific to only one time-course. We found extensive regulation of multiple Wnt pathway components in both the regenerating head and during budding. In contrast, we found upregulation of the antagonists of the TGF-β superfamily ligands only during regeneration, which suggests a coordinated downregulation of the TGF-β/BMP pathway is not necessary for budding. In addition, differential usage of chromatin remodelers and silencers during head regeneration and budding suggests that control of gene expression is crucial during these distinct developmental processes. Our results show that budding and regeneration in Hydra have distinct transcriptional profiles involving different combinations of pathways to generate a head.

2.2 Introduction

An adult Hydra polyp has a simple structure consisting of a cylindrical tube with an apical head and a basal foot. The epithelial cells of both the ectoderm and endoderm of the body column are constantly in the mitotic cycle (Campbell, 1967). As a consequence, tissue is continuously displaced towards and sloughed off at the two extremities (Campbell, 1967). Thus, the tissue dynamics of the animal involves a steady state of production and loss of tissue. To maintain the structure of an adult Hydra in this context, a small set of cells referred to as the head organizer are located in the tip of the hypostome (Broun & Bode, 2002; Browne, 1909; Technau et al., 2000). The head organizer actively maintains the pattern and morphology of the animal in the context of its tissue dynamics by signaling neighboring cells to adopt differentiated states appropriate to the head (hypostome and tentacles). When a Hydra is bisected anywhere along the body column, a head regenerates at the apical end of the lower part of the bisected animal that first involves head organizer formation (Bode, 2003, 2012).

The canonical Wnt pathway plays a critical role in a number of patterning processes in bilaterian embryos. β-catenin is involved in setting up the embryonic organizer in frog (*Xenopus)* (Guger & Gumbiner, 2000) and zebrafish (*Danio*) embryos (Schneider, Steinbeisser, Warga, & Hausen, 1996). This pathway also affects axial patterning in embryos of several species including the formation of the AP axis in mice (Haegel et al., 1995), the patterning of the AV axis in sea urchins (Logan, Miller, Ferkowicz, & McClay, 1998) as well as the establishment of the AP polarity of segments in *Drosophila* (Nüsslein-volhard & Wieschaus, 1980). The Hydra Wnt and TCF genes of the canonical Wnt pathway are expressed in the hypostome where the organizer is located (Hobmayer et al., 2000). A critical component of organizer formation is β-catenin (Gee et al., 2010). When Hydra are treated with alsterpaullone, which blocks the degradation of β-catenin by GSK3β (Leost et al., 2000), the level of β-catenin is elevated

throughout the body column and results in numerous head organizers forming all along the body colum (Broun, M., Gee L., Reinhardt B., 2005). In addition, a number of other genes have been shown to affect, or be associated with head organizer formation. These include Goosecoid (Broun, Sokol, & Bode, 1999), Brachyury (Technau & Bode, 1999), Forkhead/HNF-3b (Martinez et al., 1997), and Chordin (Rentzsch, Guder, Vocke, Hobmayer, & Holstein, 2007).

Head organizer also appears during bud formation, Hydra's asexual form of reproduction. Under normal physiological conditions Hydra reproduces asexually through budding in the lower body column. During the initial stage of bud formation, a head organizer is formed in the budding zone, and subsequently directs the formation of a bud, which eventually develops into an adult Hydra polyp. In addition to the their role in the formation and maintenance of the head organizer at the hypostome, Hydra Wnt genes are also involved in the budding process as they are expressed at the budding zone where the presumptive bud arises and in the hypostome of the growing bud (Hobmayer et al., 2000). Since both regeneration and budding involve the formation of a head organizer, a natural question is the extent to which the two gene expression programs are similar. Specifically, what are the common and regeneration-specific (or budding-specific) sets of genes involved in head organizer genesis during head regeneration and budding, and subsequently its activity and maintenance?

RNA-seq has enabled gene expression profiling, full-transcript assembly, allele-specific expression profiling and RNA-editing studies (Conesa et al., 2016). The availability of the Hydra genome (Chapman et al., 2010) has facilitated genome-wide studies of Hydra biology and enabled genome-based, *ab initio* RNA-seq of annotated and novel transcripts. During the last 5 years, RNA-seq has been used to assemble a transcriptome of Hydra (Wenger & Galliot, 2013),

to characterize the transcriptome and proteome of Hydra during head regeneration (Petersen et al., 2015), and to profile the small non-coding RNA repertoire of Hydra (Krishna et al., 2013).

In this study we used RNA-seq to characterize genome-wide gene expression patterns during Hydra head regeneration and budding. We bisected Hydra at the mid-body column, allowed them to regenerate for set periods of time, isolated the regenerating heads as well buds at various stages of growth and carried out RNA-seq (Fig. 1). We analyzed the resulting differentially expressed genes to assess the common and divergent sets of genes between head regeneration and budding in Hydra.

2.3 Results

Experimental design of Hydra head regeneration and budding time courses and body map

We performed time-course RNA-seq experiments in Hydra during head regeneration and budding as well as certain body parts. For head regeneration, animals were bisected between the regions R1 and R2 and allowed to regenerate for certain time periods (0, 2, 4, 6, 12, 24, and 48 hours) and the regenerating tips were collected for RNA sequencing (Fig. 1). For budding, the heads of developing buds at specific stages of budding (S1, S3, S4, S5, S6, S7, S8, and S10) as classified by Otto *et al.* (Otto & Campbell, 1977) were collected for RNA sequencing (Fig. 1). In addition body column, budding zone, foot, tentacle, and hypostome tissues were collected for RNA sequencing to generate a gene expression "body" map for Hydra. All experimental samples were done in two biological replicates for reproducibility. The samples libraries were built using the same protocol and the datasets were processed uniformly.

Ab initio transcriptome assembly and functional annotation of transcripts

We used the genome sequence and Augustus predicted gene models from Hydra 2.0 Genome Project (https://research.nhgri.nih.gov/hydra/). The Augustus predicted gene models consist of 33820 gene loci and 36059 transcript models. To augment the predicted gene models from the Hydra 2.0 Genome Project, we used our RNA-seq datasets to assemble *ab initio* transcripts and merge them with the existing Augustus gene models. Briefly, RNA-seq reads from all samples were mapped to the Hydra genome using STAR aligner and subsequently StringTie was used to assemble evidence-based transcript models. This approach increased the number of gene loci to 34184 and the number of transcripts to 51290.

Divergent and convergent transcriptomes of Hydra head regeneration and budding

Hydra head regeneration is a classic example of the critical role of signaling pathways in axial patterning and head organizer formation and maintenance (Bode, 2012). Head regeneration has been studied using both developmental approaches (Hobmayer et al., 2000) and more recently using genomics approaches (Petersen et al., 2015). However, less is known about the extent to which regeneration and budding gene regulatory programs overlap.

We used principal component analysis (PCA) to compare globally the regeneration and budding transcriptomes as well as the body map samples. The first two principal components (PC1 and PC2) reveal that the regeneration time points cluster separately from the budding stages and that the two processes are separated from each other along PC1 which accounts for the highest amount of variance (35%) (Fig. 2a). Such sharp clustering and separation of regeneration and budding samples along PC1 indicates that there is a large set of genes whose expression are very specific to only one of the time courses. In contrast, Principal components 2 and 3 (PC2 and PC3) reveal that the time courses of head regeneration and budding follow

slightly different trajectories to converge at a single point corresponding to the hypostome samples (Fig. 2b).

Time-series analysis of Hydra head regeneration and budding transcriptomes

The PCA analysis of Hydra head regeneration and budding shows that there are sets of genes specific or common to both time-courses. For a higher resolution analysis to define these sets of genes, we performed a time-series analysis of the regeneration and budding transcriptomes using maSigPro (Nueda, Tarazona, & Conesa, 2014) to find clusters of differentially expressed genes that have similar expression profiles. This analysis identified 5247 differentially expressed genes (6305 transcripts) that form thirteen non-redundant clusters (Fig. 3a shows all clusters with temporal profiles of some selected clusters shown in Fig. 3b. For all expression profiles of 13 clusters see Fig. 4). Ten of the clusters are specific to only one of the time courses: six clusters (cluster # 4-9) are specific to head regeneration and demonstrate complex temporal expression profiles, whereas four of the clusters (cluster # 10-13) are specific to budding (Fig. 3a). Genes in three clusters (cluster # 1-3) are expressed in both budding and head regeneration and show increasing expression along the time courses (Fig. 3a).

Gene ontology enrichment of clusters of differentially expressed transcripts

We first annotated the merged Hydra 2.0 Genome Project predicted transcripts and our *ab initio* transcripts with Gene Ontology (GO) terms using Blast2GO (Conesa et al., 2005) with 31515 of total 51290 transcripts being annotated with at least one GO term. The Genes in each of the thirteen distinct clusters (Fig. 3a) were tested for enrichment of GO terms using Blast2GO's Fisher exact test (FDR ≤ 0.05). Genes in ten clusters (cluster # 1-3, 5-7, and 10-13) were

enriched for GO terms (Fig. 3c) while three clusters (cluster # 4, 8, and 9) did not have any significantly enriched GO terms (Fig. 3c). Some representative GO terms for the clusters with enriched GO terms are shown in Fig. 3c.

Cluster 1 (192 genes, 211 transcripts), cluster 2 (128 genes, 136 transcripts), and cluster 3 (323 genes, 340 transcripts), which are upregulated in mid-to-late stages of head regeneration and budding, are enriched for GO terms related to "signaling receptor activity", "Wnt signaling pathway", and "G-protein coupled receptor signaling pathway" respectively. Cluster 5 (1434 genes, 1569 transcripts) is upregulated during early head regeneration time points (4-6 hours post-bisection) and is enriched in GO term such as "apoptotic process". Cluster 6 (256 genes, 279 transcripts), which has maximum expression at 12 hours post-bisection, was enriched in "MAP kinase phosphatase activity" and "sequestering of BMP from receptor via BMP binding". Cluster 7 (685 genes, 760 transcripts), with maximum expression at 4-6 hours post-bisection, was enriched in "receptor regulator activity" and "growth factor activity". The clusters with transcripts specific to budding stages only such as cluster 10 (572 genes, 645 transcripts), cluster 11 (280 genes, 305 transcripts), cluster 12 (229 genes, 268 transcripts), and cluster 13 (183 genes, 195 transcripts) were enriched in "ion transport" and "kinase activity", "chromatin silencing", "peptidase activity", and "DNA metabolic process" respectively.

Analysis of signaling pathways

The canonical Wnt pathway (Broun, M., Gee L., Reinhardt B., 2005; Hobmayer et al., 2000) and the MAPK pathway (Arvizu, Aguilera, & Salgado, 2006) are known to play an important role in the establishment and maintenance of the Hydra head organizer. This is indeed reflected in the number of GO terms obtained for our clusters of differentially expressed genes

(cluster # 1-3, 6, 10) related to signaling (Fig. 3c). We observed extensive regulation of canonical Wnt pathway components during the time-courses of head regeneration and budding. Several Wnt ligands (including Wnt3, Wnt1, and Wnt11) are upregulated during both head regeneration and budding (Fig. 3a). On the other hand, gremlin-2, antagonist of the TGF-β superfamily of ligands, is upregulated by 12 hours during head regeneration only, which suggests that inhibition of TGF-β superfamily signaling play an important role in head regeneration, but not during budding. This implies that the sequestration of BMP ligands to remove antagonism to the Wnt pathway is necessary during head regeneration.

2.4 Discussion

We carried out time-course experiments in Hydra using RNA-seq to obtain a comparative genome-wide view of gene expression during head regeneration and budding. We profiled 7 time points (0, 2, 4, 6, 12, 24, 48 hours) post bisection of head regeneration and 8 stages of budding (S1, S3, S4, S5, S6, S7, S8, and S10 (Otto & Campbell, 1977). 5247 genes (6305 transcripts) are differentially expressed that form thirteen clusters of unique temporal profiles along the time-courses of head regeneration and budding. Only 643 genes (687 transcripts) clustered in three distinct expression profiles are common to both head regeneration budding, while 3890 genes (4205 transcripts) and 1264 genes (1413 transcripts) are specific to head regeneration and budding respectively.

Gene ontology (GO) enrichment analysis of the clusters of differentially expressed transcripts recapitulates biological processes along the time-courses of head regeneration and budding in Hydra (Fig. 3c). This analysis reveals the biological processes common to both head regeneration and budding such as “Wnt signaling pathway”, as well as time-course specific

processes such as injury related responses during head regeneration ("apoptotic process" enriched in cluster 5 consisting of transcripts expressed maximally during early stages of head regeneration) and budding specific gene expression regulation ("chromatin silencing" enriched in cluster 13).

The role of the canonical Wnt pathway in the head organizer of hydra has extensively been documented (Broun, M., Gee L., Reinhardt B., 2005; Hobmayer et al., 2000). Many lines of evidence suggest its critical role in setting up and maintaining the hydra head organizer. *In situ* hybridization of multiple Wnt ligands has shown that their expression is restricted to the hypostome of an adult hydra, as well as to the apical tip of an evaginating bud or a regenerating head (Lengfeld et al., 2009). Furthermore, β-catenin was shown to be localized in the nuclei of epithelial cells only at the apex of the hypostome (Broun, M., Gee L., Reinhardt B., 2005). In this study we observed extensive regulation of canonical Wnt pathway components during the time-courses of head regeneration and budding, consistent with these previous studies. Seven Wnt ligands (including Wnt3, Wnt1, and Wnt11) are upregulated along the time-courses of both head regeneration and budding (Fig. 3a).

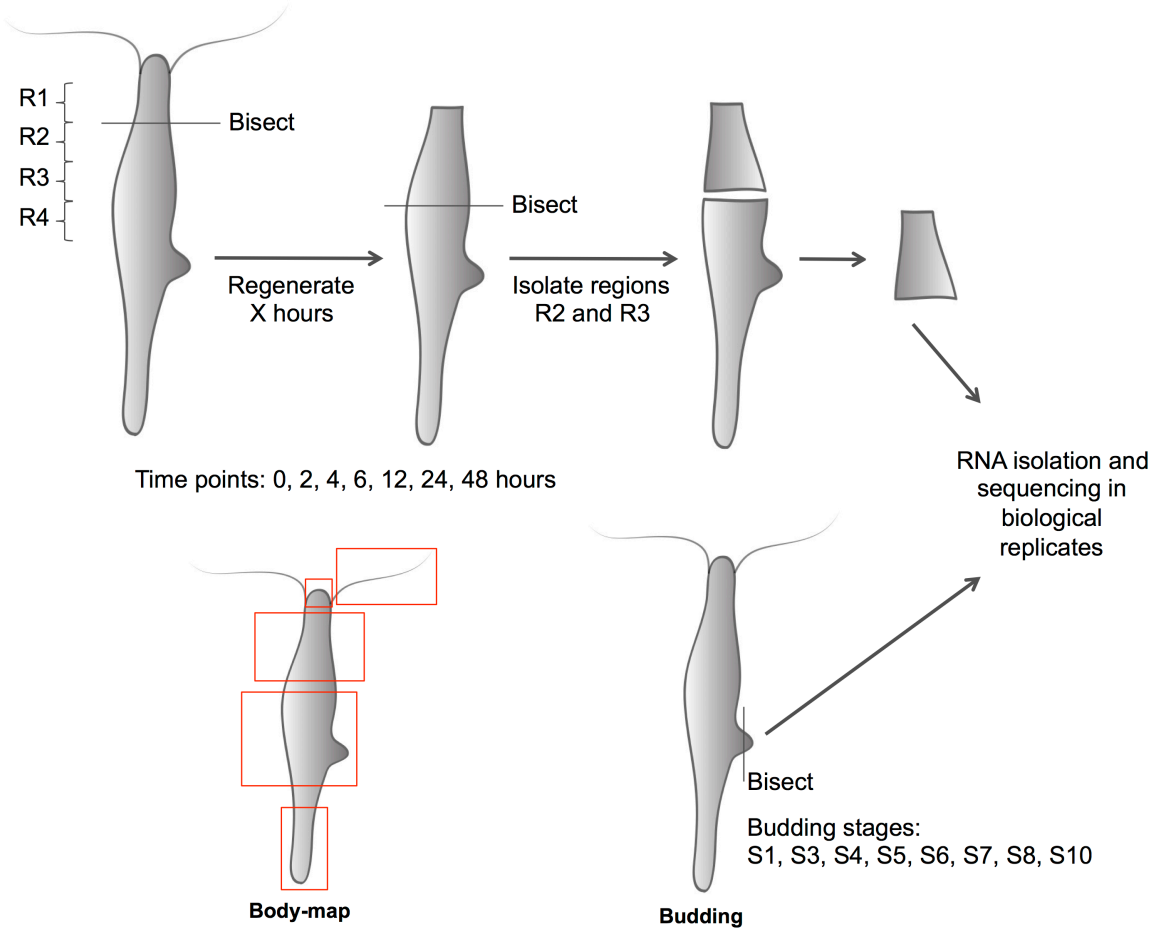
The TGF-β/SMAD pathway seems to be downregulated during head regeneration only as the GO term "sequestering of BMP from receptor via BMP binding" is enriched in cluster 6 consisting of transcripts highly expressed at 12 hours post bisection of Hydra head (Fig 3a and 3c). Downregulation of the TGF-β/SMAD signaling pathway is consistent with the idea that the role of TGF-β/SMAD signaling pathways in developing organizers in bilaterians also originated earlier in evolution as it is present in hydra (Hobmayer et al., 2004). Upregulation of gremlin-2, antagonist of the TGF-β superfamily of ligands, suggests that inhibition of TGF-β superfamily signaling occurs in the regenerating head but not during budding. Wnt and BMP pathways

exhibit mutual antagonism (Carnac et al., 1996) (Wiersdorff, Lecuit, Cohen, & Mlodzik, 1996). Early upregulation of gremlin-2 (by 12 hours) suggests that inhibition of the BMP pathway takes place to remove antagonism to the Wnt pathway during hydra head regeneration.

The time-course RNA-seq experiments in this study has shed light on genome-wide gene expression patterns during formation of the head organizer in Hydra during head regeneration and budding. The head organizer in hydra is estimated to consist of 50-300 cells at the apical tip of the head. Single animal profiling at a greater temporal resolution should provide additional insights into the establishment of head organizer in different developmental scenarios in Hydra and the processes that initiate and maintain it. Whether the head organizer plays a role in sexual embryonic development is not known, although it is likely also involved in the development of the structure of the animal during embryogenesis. Future extensions of this study to the comparison with the head organizer formation during sexual embryogenesis will reveal the extent of reuse of the normal developmental program during head regeneration.

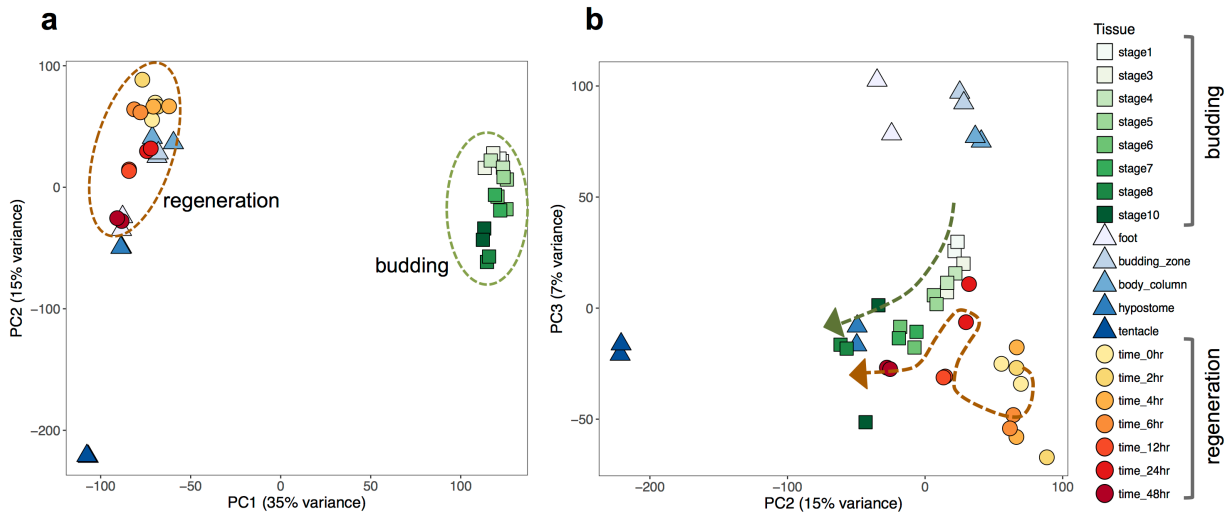
2.5 Figures

Figure 2.1: Experimental design of head regeneration and budding RNA-seq time courses.



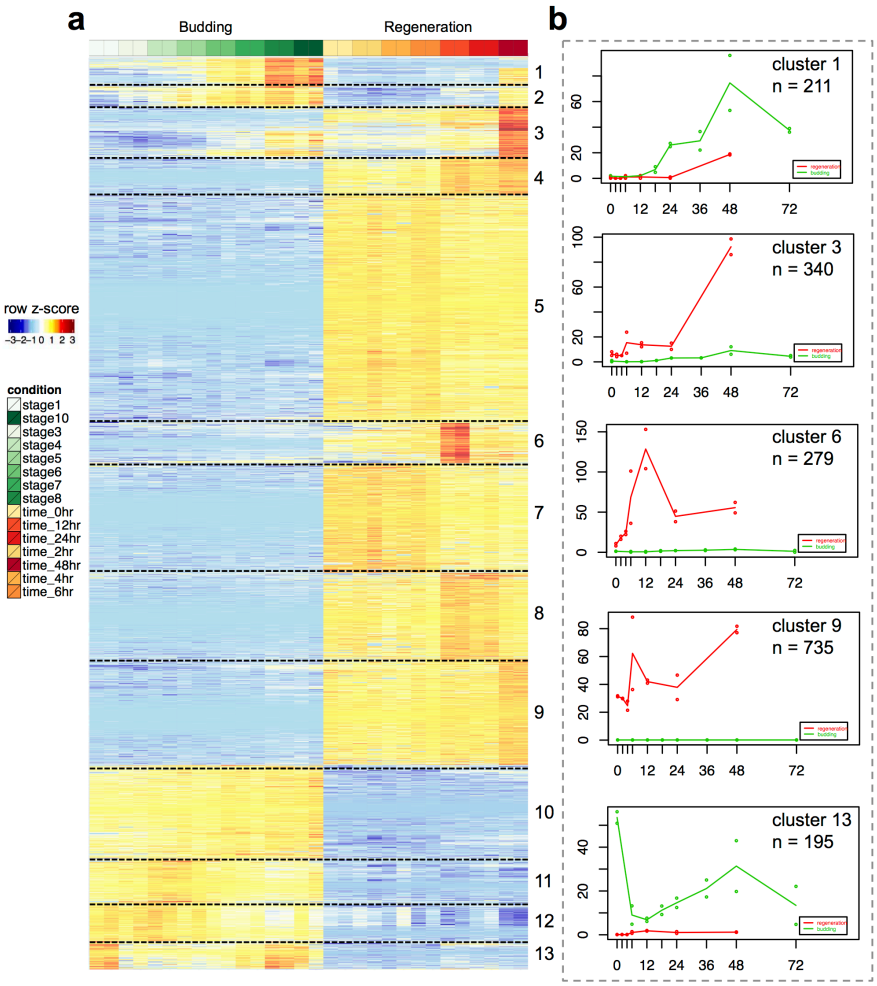
For the head regeneration experiment, Hydra were bisected at the boundary of regions R1 and R2 and allowed to regenerate for specific time periods (0, 2, 4, 6, 12, 24, and 48 hours). The regions R2 and R3 were isolated for RNA extraction and subsequent RNA-seq. For the budding experiment, the Hydra bud heads at various stages of budding (S1, S3, S4, S5, S6, S7, S8, and S10) were bisected and total RNA was extracted for RNA-seq. For "body map" tissues from tentacles, hypostome, body column, budding zone, and foot were harvested for RNA extraction.

Figure 2.2: Comparative analysis of gene expression between head regeneration and budding in Hydra using principal component analysis (PCA).



PCA analysis was performed on quantile normalized RNA-seq data for the regeneration and budding time courses as well as the various Hydra body parts. Squares denotes stages of budding, triangles the various Hydra body parts, while circles denote the time points of head regeneration post bisection. (a) The first principal component (PC1, 35% variance) reveals clear separation between the budding and regeneration time courses. (b) The second and third principal components (PC2 and PC3) reveal that the processes of regeneration and budding converge by late stages.

Figure 2.3: Clustering of differentially expressed transcripts, their temporal expression profiles and associated enriched gene ontology terms.



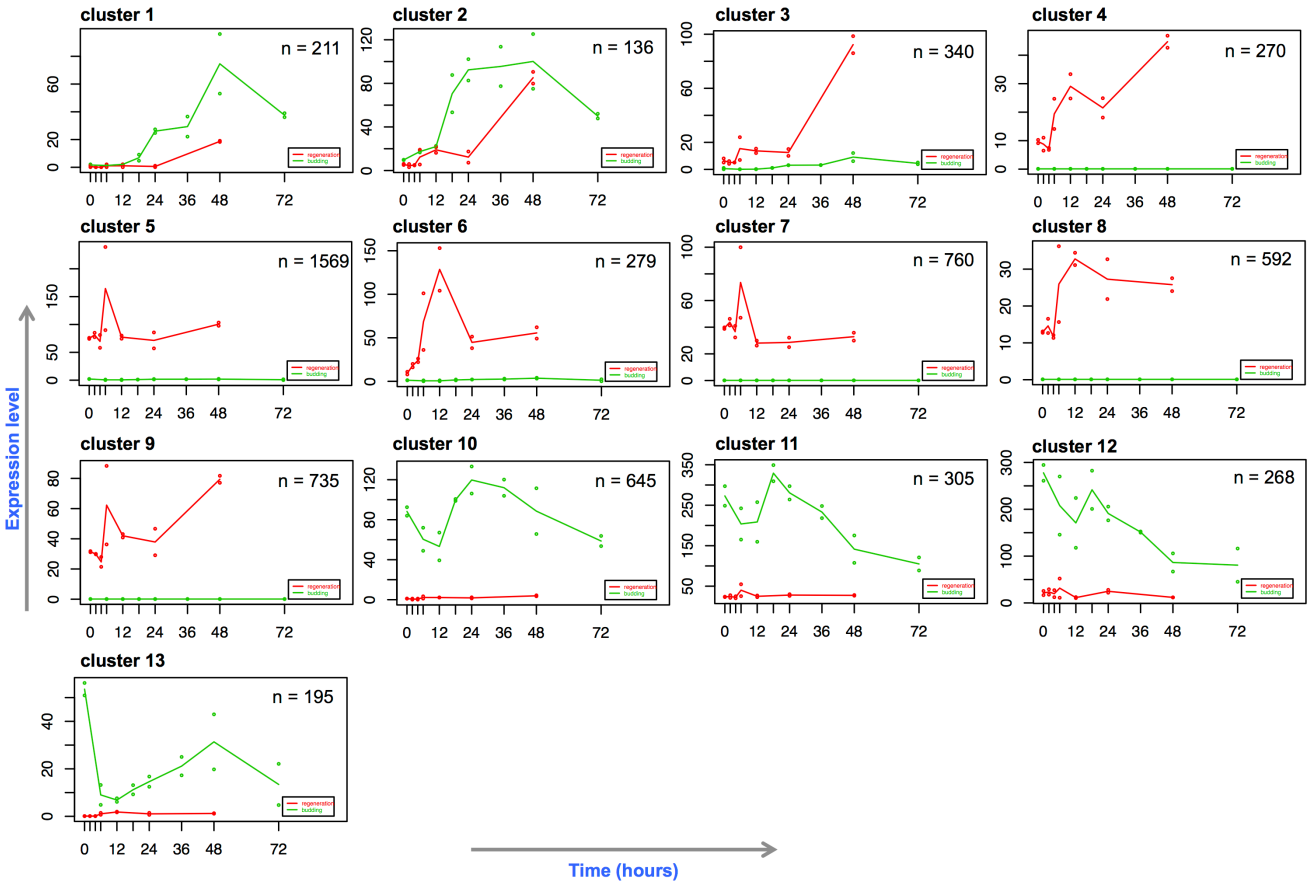
c

cluster	GO term	FDR	# test/ref
1	Signaling receptor activity	1.40E-09	24/818
2	Wnt signaling pathway	5.09E-08	8/64
2	Anatomical structure development	2.98E-05	12/563
3	G-protein coupled receptor signaling pathway	1.11E-04	21/681
5	Apoptotic process	1.82E-05	30/243
5	Regulation of apoptotic process	5.29E-04	22/173
6	MAP kinase phosphatase activity	0.017	3/4
6	Sequestering of BMP from receptor via BMP binding	0.037	2/2
7	Receptor regulator activity	9.71E-04	12/103
7	Growth factor activity	0.016	9/77
10	Ion transport	1.19E-07	43/1004
10	Kinase activity	3.43E-04	38/1186
11	Chromatin silencing	0.035	3/16
12	Peptidase activity	1.79E-04	23/1071
13	DNA metabolic process	2.36E-10	42/2695

(a) Heatmap of 6305 differentially expressed transcripts, grouped in 13 clusters based on similar expression profiles, were determined by combined time-series analysis of head regeneration and budding. The quantile-normalized transcripts per million (TPM) values of the transcripts converted to z-scores are used. The cluster numbers are indicated to the right of each cluster. (b) Representative profiles of differentially expressed transcript clusters. The cluster number corresponding to Fig. 3a and the number of transcripts in each cluster are shown inset. (c) Representative enriched GO terms for the clusters of differentially expressed genes in Fig. 3a.

Significance level is at FDR of 0.05. The numbers of transcripts with the GO terms in test set and the reference set are indicated in the fourth column.

Figure 2.4: Comparative time-series analysis of gene expression between head regeneration and budding in Hydra.



The graphs represent the median profiles of 6305 differentially expressed genes along the time courses of regeneration (red profiles) and budding (green profiles). The x and y axes represent time (hours) and gene expression levels respectively. The numbers of genes in each cluster are indicated under each graph.

2.6 Methods and Materials

Hydra culture

Hydra vulgaris polyps were used for the isolation of RNA. They were fed freshly hatched *Artemia salina* nauplii twice per week and cultured as described previously (Smith, Gee, BlitzIl, & Bode, 1999). Animals were starved for at least 1 day before any tissue manipulation or RNA isolation was carried out.

Experimental design and tissue manipulation

For each sample, 1-day starved asexual Hydra were selected. For regeneration, 1 animal per sample (with two biological replicates) was bisected at the region 1 (R1) and region 2 (R2) border (Fig.1) and allowed to undergo head regeneration for a specific period of time (0, 2, 4, 6, 12, 24, or 48 hours). Then the R2-R3 region of the animal of a sample was isolated for RNA extraction. For the budding experiment, the head region of buds from animals at various stages of budding (S1, S3, S4, S5, S6, S7, S8, or S10) (Otto & Campbell, 1977) was bisected and used for total RNA extraction (Fig. 1). Tissues from tentacles, budding zone, body column, hypostome, and foot were harvested for RNA extraction.

Total RNA extraction

Each isolated tissue was dissolved in Qiagen RNeasy buffer RLT (with 2-betamercaptoethanol added) within 3 minutes of isolation. The dissolved tissue was immediately used for total RNA isolation using Qiagen RNeasy kit according to manufacturer's protocol. The total RNA for each sample was treated with DNase from TURBO DNA-free kit to remove any

genomic DNA contamination. RNA quality was checked with Agilent Bioanalyzer and samples with RIN scores ≥9 were used for RNA-seq library preparation.

Illumina library preparation

Mulitplexed RNA-seq libraries were built using the Smart-seq2 protocol (Picelli et al., 2014) with slight modifications. Briefly, mRNA from total RNA in each sample was converted to full-length cDNA using poly-dT primer and reverse transcriptase. cDNA was amplified using appropriate number of PCR cycles based on the initial amount of total RNA and as recommended by the Smart-seq2 protocol. 20 ng full-length cDNA for each sample was converted to sequencing library by tagmentation with the Illumina Nextera kit. 8 cycles of PCR were used for library amplification. Libraries were multiplexed and sequenced as 43 bp Illumina paired-end reads.

***Ab initio* transcriptome assembly and functional annotation of transcripts**

We used the genome sequence and Augustus predicted gene models from Hydra 2.0 Genome Project (https://research.nhgri.nih.gov/hydra/) for the transcriptome assembly and gene expression analysis. Adapter sequences and low quality base pairs from the paired-ends reads were trimmed using Trimmomatic v. 0.35 (Bolger, Lohse, & Usadel, 2014) using the following parameters: "PE [read1.fastq] [read2.fastq] pe_read1.fastq.gz se_read1.fastq.gz pe_read2.fastq.gz se_read2.fastq.gz ILLUMINACLIP:NexteraPE-PE.fa:2:30:8:4:true LEADING:20 TRAILING:20 SLIDINGWINDOW:4:17 MINLEN:30". The trimmed reads were mapped to the Hydra genome using STAR v. 2.4.2a (Dobin et al., 2013) using the following parameters: "--outFilterMultimapNmax 20 --alignSJoverhangMin 8 --alignSJDBoverhangMin 1 --

outFilterMismatchNmax 999 --outFilterMismatchNoverReadLmax 0.04 --alignIntronMin 20 --alignIntronMax 1000000 --alignMatesGapMax 1000000 --outSAMunmapped Within --outFilterType BySJout --outSAMattributes NH HI AS NM MD XS --outSAMstrandField intronMotif --outSAMtype BAM SortedByCoordinate --sjdbScore 1". The mapped reads from the two biological replicates from each sample were pooled and *ab initio* transcripts were assembled using StringTie v. 1.3.4b (Pertea et al., 2015) using the following parameters: "-G [GTF file] -o stringtie.gtf -c 3 -p 12 –A stringtie.abundance.txt". The assembled transcripts for all samples were merged with the Hydra 2.0 Genome Project Augustus predicted models to obtain a final reference transcriptome.

The reference transcriptome was annotated with GO terms using Blast2GO (Conesa et al., 2005). First, a BLAST search was done for all the transcripts against NCBI's non-redundant NR database. The transcripts were then annotated with the GO terms associated with the BLAST hits using the "Mapping" and "Annotation" functions of Blast2GO. The GO terms were expanded using the InterProScan and Annex mapping utilities of Blast2GO.

Gene expression analysis

RNA-seq reads for each sample were mapped to the reference transcriptome using Bowtie v. 1.2 (Langmead, Trapnell, Pop, & Salzberg, 2009) with the following parameters: "-X 2000 -a -m 200 -S --seedlen 25 -n 2 -v 3". Transcript expression levels and read counts were obtained using RSEM v. 1.2.31 (Li & Dewey, 2011) with the following parameters: "--paired-end --num-threads 8 --calc-ci".

Time-series analysis of budding and head regeneration time courses was done using maSigPro v. 1.42.0 (Nueda et al., 2014) and R v. 3.2.3 using the maSigPro functions "p.vector"

and “T.fit” with a significance level of 0.01 to find clusters of differentially expressed transcripts and their temporal dynamics.

The heatmap of differentially expressed transcripts was generated using R. The transcripts per millions (TPM) values of the differentially expressed transcripts were log2 transformed and scaled for generating the heatmap.

Gene ontology (GO) analysis

Each maSigPro cluster of differentially expressed genes (Fig. 3a) was analyzed for GO enrichment using the Fisher’s exact test function of Blast2GO. Each cluster was tested for GO enrichment using the entire reference transcriptome as the reference set. FDR of 5% was used as the significance threshold.

2.7 References

Arvizu, F., Aguilera, A., & Salgado, L. M. (2006). Activities of the protein kinases STK, PI3K, MEK, and ERK are required for the development of the head organizer in Hydra magnipapillata. *Differentiation*, *74*(6), 305–312. https://doi.org/10.1111/j.1432-0436.2006.00078.x

Bode, H. R. (2003). Head regeneration in Hydra. *Developmental Dynamics*, *226*(2), 225–236. https://doi.org/10.1002/dvdy.10225

Bode, H. R. (2012). The head organizer in Hydra. *International Journal of Developmental Biology*, *56*(6–8), 473–478. https://doi.org/10.1387/ijdb.113448hb

Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics*, *30*(15), 2114–2120. https://doi.org/10.1093/bioinformatics/btu170

Broun, M., Gee L., Reinhardt B., B. H. (2005). Formation of the head organizer in hydra involves the canonical Wnt pathway. *Development*, *132*(12), 2907–2916. https://doi.org/10.1242/dev.01848

Broun, M., & Bode, H. R. (2002). Charcaterization of the head organzer in hydra. *Development*, *129*, 875–884.

Broun, M., Sokol, S., & Bode, H. R. (1999). Cngsc, a homologue of goosecoid, participates in the patterning of the head, and is expressed in the organizer region of Hydra. *Development (Cambridge, England)*, *126*(23), 5245–5254.

Browne, E. N. (1909). The production of new hydranths in Hydra by the insertion of small grafts. *Journal of Experimental Zoology*, *7*(1), 1–23. https://doi.org/10.1002/jez.1400070102

Campbell, R. D. (1967). Tissue dynamics of steady state growth in Hydra littoralis. II. Patterns of tissue movement. *Journal of Morphology*, *121*(1), 19–28. https://doi.org/10.1002/jmor.1051210103

Carnac, G., Kodjabachian, L., Gurdon, J. B., Lemaire, P., Cho, K. W. Y., Blumberg, B., … Rutishauser, U. (1996). The homeobox gene Siamois is a target of the Wnt dorsalisation pathway and triggers organiser activity in the absence of mesoderm. *Development (Cambridge, England)*, *122*(10), 3055–65. https://doi.org/10.1016/0092-8674(91)90288-a

Chapman, J. A., Kirkness, E. F., Simakov, O., Hampson, S. E., Mitros, T., Weinmaier, T., … Steele, R. E. (2010). The dynamic genome of Hydra. *Nature*, *464*(7288), 592–596. https://doi.org/10.1038/nature08830

Conesa, A., Götz, S., García-Gómez, J. M., Terol, J., Talón, M., & Robles, M. (2005). Blast2GO: A universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics*, *21*(18), 3674–3676. https://doi.org/10.1093/bioinformatics/bti610

Conesa, A., Madrigal, P., Tarazona, S., Gomez-Cabrero, D., Cervera, A., McPherson, A., … Mortazavi, A. (2016). A survey of best practices for RNA-seq data analysis. *Genome Biology*. https://doi.org/10.1186/s13059-016-0881-8

Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., … Gingeras, T. R. (2013). STAR: Ultrafast universal RNA-seq aligner. *Bioinformatics*, *29*(1), 15–21. https://doi.org/10.1093/bioinformatics/bts635

Gee, L., Hartig, J., Law, L., Wittlieb, J., Khalturin, K., Bosch, T. C. G., & Bode, H. R. (2010). β-catenin plays a central role in setting up the head organizer in hydra. *Developmental Biology*, *340*(1), 116–124. https://doi.org/10.1016/j.ydbio.2009.12.036

Guger, K. A., & Gumbiner, B. M. (2000). A mode of regulation of β-catenin signaling activity in Xenopus embryos independent of its levels. *Developmental Biology*, *223*(2), 441–448. https://doi.org/10.1006/dbio.2000.9770

Haegel, H., Larue, L., Ohsugi, M., Fedorov, L., Herrenknecht, K., & Kemler, R. (1995). Lack of β-catenin affects mouse development at gastrulation. *Development*, *121*(11), 3529–3537. Retrieved from http://www.scopus.com/inward/record.url?eid=2-s2.0-0028971954&partnerID=tZOtx3y1

Hobmayer, B., Rentzsch, F., Kuhn, K., Happel, C. M., von Laue, C. C., Snyder, P., … Holstein, T. W. (2000). WNT signalling molecules act in axis formation in the diploblastic metazoan Hydra. *Nature*, *407*(6801), 186–189. https://doi.org/10.1038/35025063

Krishna, S., Nair, A., Cheedipudi, S., Poduval, D., Dhawan, J., Palakodeti, D., & Ghanekar, Y. (2013). Deep sequencing reveals unique small RNA repertoire that is regulated during head regeneration in Hydra magnipapillata. *Nucleic Acids Research*, *41*(1), 599–616. https://doi.org/10.1093/nar/gks1020

Langmead, B., Trapnell, C., Pop, M., & Salzberg, S. L. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biology*, *10*(3). https://doi.org/10.1186/gb-2009-10-3-r25

Lengfeld, T., Watanabe, H., Simakov, O., Lindgens, D., Gee, L., Law, L., … Holstein, T. W. (2009). Multiple Wnts are involved in Hydra organizer formation and regeneration. *Developmental Biology*, *330*(1), 186–199. https://doi.org/10.1016/j.ydbio.2009.02.004

Leost, M., Schultz, C., Link, a, Wu, Y. Z., Biernat, J., Mandelkow, E. M., … Meijer, L. (2000). Paullones are potent inhibitors of glycogen synthase kinase-3beta and cyclin-dependent kinase 5/p25. *European Journal of Biochemistry / FEBS*, *267*(19), 5983–5994. https://doi.org/10.1046/j.1432-1327.2000.01673.x

Li, B., & Dewey, C. N. (2011). RSEM: Accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics*, *12*. https://doi.org/10.1186/1471-2105-12-323

Logan, C. Y., Miller, J. R., Ferkowicz, M. J., & McClay, D. R. (1998). Nuclear beta-catenin is required to specify vegetal cell fates in the sea urchin embryo. *Development (Cambridge, England)*, *126*(2), 345–357. Retrieved from http://dev.biologists.org/content/126/2/345.long%5Cnpapers://e4e91285-7e3b-4e20-9d75-546bc15a52f8/Paper/p1041

Martinez, D. E., Dirksen, M. L., Bode, P. M., Jamrich, M., Steele, R. E., & Bode, H. R. (1997). Budhead, a fork head/HNF-3 homologue, is expressed during axis formation and head specification in hydra. *Developmental Biology*, *192*(2), 523–536. https://doi.org/10.1006/dbio.1997.8715

Nueda, M. J., Tarazona, S., & Conesa, A. (2014). Next maSigPro: Updating maSigPro bioconductor package for RNA-seq time series. *Bioinformatics*, *30*(18), 2598–2602. https://doi.org/10.1093/bioinformatics/btu333

Nüsslein-volhard, C., & Wieschaus, E. (1980). Mutations affecting segment number and polarity in drosophila. *Nature*, *287*(5785), 795–801. https://doi.org/10.1038/287795a0

Otto, J. J., & Campbell, R. D. (1977). Budding in Hydra attenuata: Bud stages and fate map. *Journal of Experimental Zoology*, *200*(3), 417–428. https://doi.org/10.1002/jez.1402000311

Pertea, M., Pertea, G. M., Antonescu, C. M., Chang, T. C., Mendell, J. T., & Salzberg, S. L. (2015). StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nature Biotechnology*, *33*(3), 290–295. https://doi.org/10.1038/nbt.3122

Petersen, H. O., Hï¿½ger, S. K., Looso, M., Lengfeld, T., Kuhn, A., Warnken, U., … Holstein, T. W. (2015). A comprehensive transcriptomic and proteomic analysis of hydra head regeneration. *Molecular Biology and Evolution*, *32*(8), 1928–1947. https://doi.org/10.1093/molbev/msv079

Picelli, S., Faridani, O. R., Björklund, Å. K., Winberg, G., Sagasser, S., & Sandberg, R. (2014). Full-length RNA-seq from single cells using Smart-seq2. *Nature Protocols*, *9*(1), 171–181. https://doi.org/10.1038/nprot.2014.006

Rentzsch, F., Guder, C., Vocke, D., Hobmayer, B., & Holstein, T. W. (2007). An ancient chordin-like gene in organizer formation of Hydra. *Proceedings of the National Academy of Sciences*, *104*(9), 3249–3254. https://doi.org/10.1073/pnas.0604501104

Schneider, S., Steinbeisser, H., Warga, R. M., & Hausen, P. (1996). β-catenin translocation into nuclei demarcates the dorsalizing centers in frog and fish embryos. *Mechanisms of Development*, *57*(2), 191–198. https://doi.org/10.1016/0925-4773(96)00546-1

Signaling molecules in the. (2004), *3*(June), 107–114.

Smith, K. M., Gee, L., BlitzIl, & Bode, H. R. (1999). CnOtx, a mamber of the OTX gene family, has a role in cell movement in Hydra. *Developmental Biology*, *212*, 392–404.

Technau, U., & Bode, H. R. (1999). HyBra1, a Brachyury homologue, acts during head formation in Hydra. *Development (Cambridge, England)*, *126*(5), 999–1010.

Technau, U., Cramer von Laue, C., Rentzsch, F., Luft, S., Hobmayer, B., Bode, H. R., & Holstein, T. W. (2000). Parameters of self-organization in Hydra aggregates. *Proceedings of the National Academy of Sciences*, *97*(22), 12127–12131. https://doi.org/10.1073/pnas.97.22.12127

Wenger, Y., & Galliot, B. (2013). RNAseq versus genome-predicted transcriptomes: A large population of novel transcripts identified in an Illumina-454 Hydra transcriptome. *BMC Genomics*, *14*(1). https://doi.org/10.1186/1471-2164-14-204

Wiersdorff, V., Lecuit, T., Cohen, S. M., & Mlodzik, M. (1996). Mad acts downstream of Dpp receptors, revealing a differential requirement for dpp signaling in initiation and propagation of morphogenesis in the Drosophila eye. *Development (Cambridge, England)*, *122*(7), 2153–62. Retrieved from http://dev.biologists.org/content/122/7/2153.abstract

Chapter 3

The open-chromatin landscape of Hydra during head regeneration

Statement of contribution

In this study, I designed (in agreement with my thesis advisor Dr. Ali Mortazavi) and performed experiments, analyzed the data and interpreted the results.

3.1 Abstract

Cnidarians and bilaterians diverged about 600 millions years ago but the gene contents of species of both groups are surprisingly similar despite divergent morphologies and functions. Little is known about the role of *cis*-regulatory elements outside of bilaterians. Understanding gene regulatory mechanisms of cnidarians can potentially shed light on metazoan evolution. In this study we mapped the *cis*-regulatory landscape of Hydra, a well-studied cnidarian model organism, in the context of a developmental process by mapping 27,137 open-chromatin elements in its genome. We used ChIP-seq from several histone modifications to 9998 candidate promoter and 3018 candidate enhancer-like elements respectively. We show that a subset of these regulatory elements is dynamically remodeled during head regeneration. Our results show that Hydra displays complex gene regulatory structures of developmentally dynamic enhancers, which would date their evolution to predate the split of cniderians and bilaterians.

3.2 Introduction

Hydra belongs to the phylum Cnidaria that consists of ~10,000 species divided into two major groups: Anthozoa (comprising of sea anemones, corals, and sea pens) and Medusozoa (sea wasps, jellyfish, and Hydra). The hallmark traits of cnidarians are their external radial symmetry and nematocytes, stinging cells used for predation. Unlike the more common bilaterians that include all vertebrates and most invertebrates with left-right symmetry, cnidarians consist of two germ layers (endoderm and ectoderm) and have a single body axis called the oral-aboral axis. From a phylogenetic perspective cnidarians and bilaterians diverged ~600 millions years ago (Technau & Steele, 2012). Therefore the study of cnidarians provides potential opportunities for elucidating key aspects of metazoan evolution such as the formation of mesoderm, bilaterian body plan and the nervous system. Considering the important evolutionary insights that can be

obtained from comparison of cnidarians and bilaterians, genome sequencing and functional genomic studies of cnidarians have been at the forefront. Such efforts culminated in the sequencing of the genomes of the anthozoan *Nematostella vectensis* (Putnam et al., 2007) in 2007 and of the medusozoan *Hydra vulgaris* (Chapman et al., 2010) in 2010. A rather surprising finding of the genome sequencing of Hydra and *Nematostella* was that the gene contents of these basal metazoans are similar to those of bilaterians (Chapman et al., 2010; Putnam et al., 2007). This finding led to speculation that the difference in the body plans of cnidarians and bilaterians is due to differences in gene regulation (Schwaiger, 2014) based on findings that body plan evolution is often a consequence of changes in gene regulation (Carroll, 2008) and differences in *cis*-regulatory elements among even closely related species (Frankel et al., 2011; Villar et al., 2015).

Gene expression at the transcriptional level is controlled by DNA sequences called *cis*-regulatory modules (CRMs). CRMs are short genomic regions that play specific roles in controlling the expression of their associated genes. CRMs act as binding sites for sequence-specific as well as general transcription factors (TFs) that mediate the expression of genes. CRMs are classified into promoters and enhancers. Promoters are short genomic regions overlapping the transcription start sites (TSSs) of genes and act as binding sites for general transcription factors that complex with RNA Polymerase II (Maston, Evans, & Green, 2006). Enhancers are *cis*-acting 200-1000 base pair (bp) DNA segments that contain multiple transcription factor binding sites (TFBSs) and control the expression of target genes (Levine, 2010). They are sometimes located tens to hundreds of kilobases (Kb) away from their associated genes (Levine, 2010) and can be located up to 1 megabase (Mb) from their targets (Lettice et al., 2003). Once the right combination of transcription factors are bound to the TFBSs

at an enhancer, the enhancer region is brought within close proximity of the target gene promoter by DNA looping (Bulger & Groudine, 1999).

A comparison of the gene regulatory landscapes of genomes in the cnidarians and bilaterians would require systematic genome-wide mapping of *cis*-regulatory elements within sequenced cnidarian genomes. So far, this has only been attempted in *Nematostella*, leading to identification of over 5000 enhancers and the surprising finding that the gene regulatory landscape of *Nematostella* is at least as complex than those of bilaterians (Schwaiger, 2014). Although studies in mammalian model systems show that enhancers evolve at a more rapid pace than the coding sequences, it remains to be determined if enhancer evolution is an important agent of metazoan evolution over long evolutionary periods.

The first step in a comparative study of enhancers in cnidarians is a genome-wide identification of enhancers and other regulatory elements. While bioinformatics approaches have been used to locate enhancers genome-wide, the most reliable approach involves experimental methods to identify genomic regions containing features associated with enhancers. Experimental methods for mapping promoters and enhancers are based on detecting specific histone modifications that are associated with each CRM type as well as increased accessibility of such regions due to localized depletion of nucleosomes. Specific combinations of posttranslational modifications of histone proteins are associated with either promoters or enhancers. For example, histone H3 proteins in promoter regions are marked by high levels of trimethylation or dimethylation at Lys 4 (H3K4me3 and H3K4me2, respectively) in both fungi, plants, and animals while high levels of H3K4me2 (low levels of H3K4me3) and H3K27ac (H3 Lys 27 acetylation) mark active enhancer regions in bilaterians. Chromatin immunoprecipitation (ChIP-seq) with an antibody recognizing a specific histone modification followed by sequencing

can be used to map the locations of active promoters and enhancers genome-wide. Active promoters and enhancers are located in genomic regions depleted of nucleosomes so that their TFBSs can be accessed by transcription factors. Open chromatin regions are DNA segments that are nucleosome-poor and are therefore prone to quick enzymatic digestion due to easy accessibility compared to tightly wound chromatin. This is the basis for high-throughput assays, such as DNase-seq (Boyle et al., 2011; Hesselberth et al., 2009; Neph et al., 2012) and ATAC-seq (Buenrostro, Giresi, Zaba, Chang, & Greenleaf, 2013), which preferentially digest open-chromatin regions with enzymes DNase and Tn5 transposase respectively when treated for a short period of time. ATAC-seq has gained popularity over DNase-seq due to ease of protocol and low cell number requirement (500-50000 cells).

In this study we profiled the open-chromatin elements of Hydra during a 48-hour time-course of head regeneration and body map as well as corresponding datasets of three histone modifications (H3K4me2, H3K4me3, and H3K27ac) to generate genome-wide maps of candidate promoter and enhancer-like elements in Hydra. The integrative analysis of these datasets allowed us to predict sets of 9998 candidate promoters and 3018 candidate enhancer-like elements in the Hydra genome. We find evidence for extensive chromatin remodeling of the regenerating head tissue, with 17% of the promoters and 29% of the enhancers changing their chromatin state during head regeneration.

3.3 Results

Experimental design of time-course head regeneration and "body map" to map the regulatory regions in Hydra genome

We performed time-course experiments in Hydra during head regeneration and several body parts to generate a genome-wide map of regulatory elements for Hydra using assay for

transposase-accessible chromatin using sequencing (ATAC-seq) (Buenrostro et al., 2013) and chromatin immunoprecipitation of H3K4me3, H3K4me2, and H3K27ac using a modified ChIP-seq protocol (ChIPmentation) that requires lower input than the original ChIP-seq protocol (Schmidl, Rendeiro, Sheffield, & Bock, 2015) (Fig 3.1). For head regeneration, animals were bisected between regions R1 and R2 and allowed to regenerate for fixed time periods (0, 2, 4, 6, 12, 24, and 48 hours). The regenerating tips were collected to isolate nuclei and treated with Tn5 transposase to map transposase-accessible chromatin (ATAC-seq method) or the tissues were crosslinked with formaldehyde to covalently attach proteins to their genome targets (ChIP-seq method) (Fig. 3.1). In addition body column, budding zone, foot, tentacle, and hypostome tissues were collected to generate a "body map" of open-chromatin elements for Hydra. All experimental samples were done in two biological replicates for reproducibility.

Mapping the open-chromatin landscape of Hydra

Following sequencing of ATAC-seq libraries to an average depth of 20 million reads and mapping onto the latest release of the Hydra genome (Hydra 2.0 Genome Project), regions of open-chromatin corresponding to higher signal than the surrounding regions ("peaks") were called on each dataset using Homer (Heinz et al., 2010). Calls from each biological replicate were compared and only those that overlapped were retained for further analysis. The sets of peaks from all samples in the regeneration time-course and body-map were merged to obtain a consolidated set of 27,137 peaks.

We classified the 27,137 open-chromatin elements according to their genomic locations with respect to the annotated transcripts and the *ab initio* transcripts assembled from RNA-seq data (chapter 2) (Fig 3.3a). Since no prior knowledge about the locations of regulatory elements

in Hydra was available, we defined four classes of open-chromatin elements as follows: TSS (peaks within ±2 Kb of transcript start sites), intergenic (peaks located between transcripts), intronic (peaks overlapping annotated introns), and exonic (peaks overlapping annotated exons) (Fig. 3.3a). Using this classification scheme, we identified 9998 TSS open elements, 8962 intergenic open elements, 6454 intronic open elements, and 1723 exonic open elements (Fig 3.3b). We next looked at the distribution of ATAC-seq signal at the four types of open-chromatin elements defined above (Fig. 3.3c). The TSS open elements possess the highest amount of signal followed by the intergenic open elements. Most of the signal is located at the centers of the four types of peaks. Based on the genomic locations of the peaks and the enrichment of ATAC-seq signal at them and their distance from annotated TSS, our set of open-chromatin elements provides candidate promoter (near TSS) and enhancer-like elements (intergenic). However, we cannot exclude that some intergenic elements represent unannotated TSSs without using additional evidence such as the ratio of H3K4me3 to H3K4me2 from the ChIP signal as described below.

Classification of open-chromatin elements using histone modifications

Using our ATAC-seq datasets, we obtained a set of 27,137 open-chromatin elements that were classified into four groups based on their genomic locations (Fig. 3.3b). We used chromatin immunoprecipitation followed by sequencing (ChIP-seq) (Schmidl et al., 2015) in several of the corresponding time points of head regeneration time-course and body parts as well as whole animal (Fig 3.2b) to generate histone modification profiles. We used antibodies against H3 dimethylation of Lys 4 (H3K4me2), trimethylation of Lys 4 (H3K4me3), and acetylation of Lys 27 (H3K27ac) (Fig 3.2b). The histone modifications H3K4me3 and H3K4me2 are known to

mark chromatin at the promoter regions with a higher ratio of H3K4me3 to H3K4me2, whereas high H3K4me2 with low H3K4me3 predominantly marks the enhancer regions and H3K27ac marks active regulatory regions (Fig. 3.2a).

We computed and compared the normalized enrichment of the H3K4me2 and H3K4me3 at the peaks from the four sets of open-chromatin elements classified based on their genomic locations (Fig. 3.4). The TSS open-chromatin elements showed the highest enrichment of H3K4me2 and H3K4me3, with a higher enrichment of H3K4me3 (Fig. 3.4a). Thus, the chromatin marks provide further evidence for the TSS open-chromatin elements as candidate promoter regions in the Hydra genome. We expected higher enrichment of H3K4me2 compared to H3K4me3 at the remaining three classes of open-chromatin elements (intergenic, intronic, and exonic) since these were non-TSS overlapping. We observed equal (intergenic, Fig. 3.4b) or slightly lower enrichment of H3K4me2 (intronic and exonic, Fig. 3.4c-d) at these regions. A reason for the discrepancy in the relative enrichments of H3K4me2 and H3K4me3 at the non-TSS open-chromatin element sets could be the inclusion of peaks overlapping the non-annotated TSS regions. Therefore, we used the relative enrichment of H3K4me2 over H3K4me3 to score the peaks (Fig 3.5) and identify candidate enhancer regions. We defined H3K4me2 enriched peaks as having 50% or higher enrichment of H3K4me2 signal over H3K4me3 (Fig 3.5a). This strategy led to the identification of 3018 ATAC-seq peaks which are predominantly intergenic (1918/3018) followed by intronic (853/3018) and exonic open-chromatin elements (247/3018) (Fig. 3.5b). Comparison of the histone mark signals at the 3018 peaks reveals considerable enrichment of H3K4me2 relative to the H3K4me3 mark (Fig 3.5c). Therefore, the set of 3018 open-chromatin elements, based on their genomic locations and the enrichment of histone modifications, form the likeliest candidates for enhancer-like regions in the Hydra genome.

Dynamics of open-chromatin elements during Hydra head regeneration

Hydra head regeneration is a dynamic process involving changes in expression of multiple genes related to Wnt signaling pathway (Lengfeld et al., 2009), MAPK pathway (Arvizu, Aguilera, & Salgado, 2006), and response to injury (Petersen et al., 2015) to name a few. An important question that remains unanswered is: How extensive is the remodeling of chromatin in Hydra genome in response to bisection and regeneration of head? With a genome-wide set of open-chromatin elements obtained from data generated in this study, we explored the above question.

Dynamic remodeling of the chromatin during Hydra head regeneration time-course can be observed at the Wnt3 gene locus (Fig. 3.6) which is known to be one of the earliest Wnt ligands expressed at the regenerating head (Lengfeld et al., 2009). Open-chromatin signals appear at the Wnt3 promoter and upstream candidate enhancer-like regions as early as 4 hours post bisection (Fig 3.6). We extended this analysis to the complete set of 27,137 open-chromatin elements by looking for differentially hypersensitive (DHS) elements genome-wide. Differential analysis of all the peaks identified 4,168 DHS open-chromatin elements, at 5% FDR (false discovery rate) and minimum 2-fold change, that form eight groups with distinct dynamic patterns when clustered (Fig 3.7a). The clusters reveal sets of open-chromatin elements specific to certain tissues or head regeneration time-course. For example, cluster 1 consists of open-chromatin elements specific to the foot, budding zone, and body column sections of Hydra, while cluster 3 and 8 consist of elements that lose or gain hypersensitivity during head regeneration respectively (Fig. 3.7a).

To further look at the candidate promoter and high-confidence enhancer-like elements, we hierarchically clustered the sets of such elements separately (Fig. 3.7b-c). There are 1712 promoter and 882 candidate enhancer elements that are DHS (Fig. 3.7b-c).

2.4 Discussion

Functional genomic analyses of cnidarians have been spurred by sequencing of the genomes of the two most widely used cnidarian model organisms: *Hydra vulgaris* and *Nematostella vectensis* (Chapman et al., 2010; Putnam et al., 2007). The genomes of the two cnidarians are similar in terms of gene content to the vertebrate model organisms (Chapman et al., 2010; Putnam et al., 2007) despite the fact that cnidarians and bilaterians diverged ~600 million years ago. A more elaborate gene regulatory system in bilaterians has been suggested for the increasing complexity in morphologies and functions in bilaterians when compared to "simpler" organisms such as plants and fungi as well as cniderians. This hypothesis can be tested by comparing the gene regulatory elements and their functions among cnidarians and bilaterians, but such elements have been studied almost exclusively in bilaterian model organisms (ENCODE Consortium et al., 2012; Roy et al., 2010; Visel et al., 2013). Recently, the gene regulatory landscape of the cnidarian *Nematostella vectensis* was studied by examining its genome-wide histone modification landscape (Schwaiger, 2014). This study reported the presence of over 5000 enhancers in *Nematostella vectensis* genome and a similar gene regulatory mechanism among bilaterians and *Nematostella* (Schwaiger, 2014). But a complete picture of the *cis*-regulatory elements and their dynamic remodeling during a developmental process in a cnidarian, such as head regeneration, is still missing.

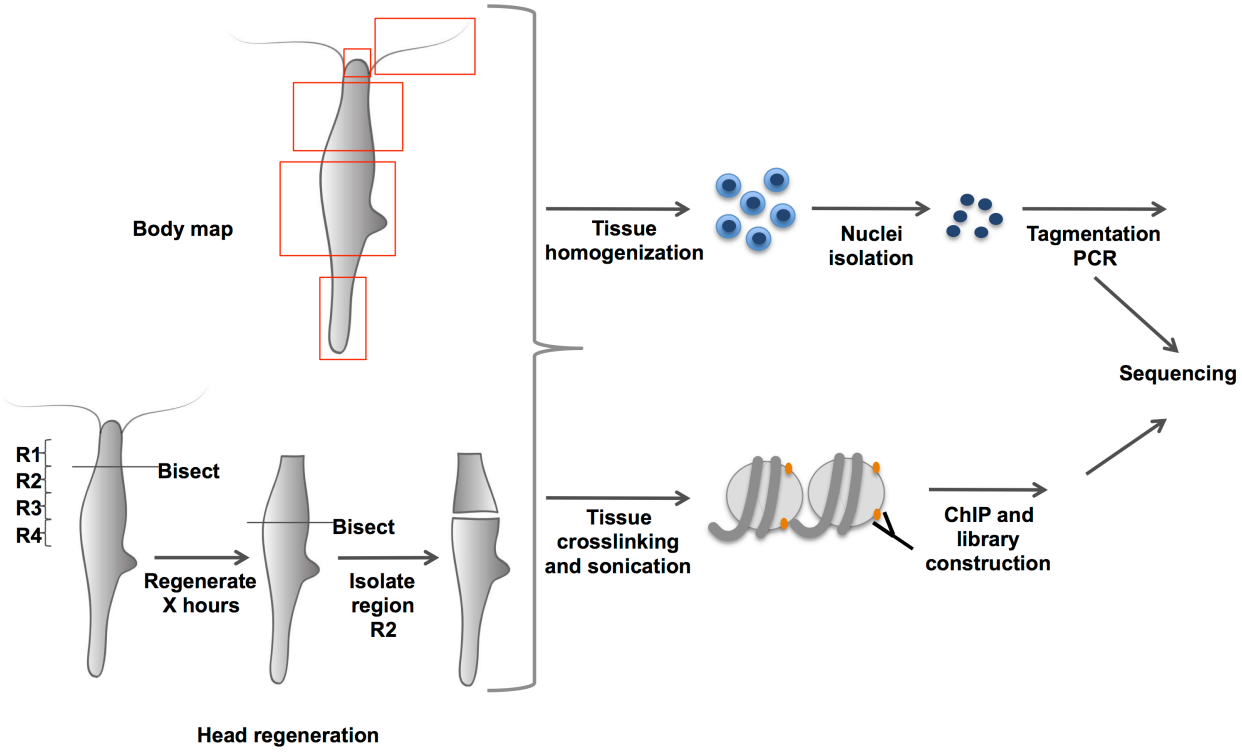
When a Hydra polyp is bisected anywhere along the body column, a head regenerates at the apical end of the lower part of the bisected animal. This process, known as head regeneration, has been a popular developmental model in understanding the role of signaling pathways in axial patterning (Bode, 2012). Thus, developmental and genomic approaches have been used to elucidate the transcriptional dynamics of this process. In this study, we provide a resource of candidate regulatory elements that change dynamically in the context of head regeneration and add a new dimension to understanding how the transcriptional changes are associated with changes at the chromatin level. We carried out a body-map and head regeneration time-course experiment in Hydra using ATAC-seq to obtain a genome-wide view of open-chromatin landscape and to identify candidate gene regulatory elements in the Hydra genome. We profiled seven time points (0, 2, 4, 6, 12, 24, 48 hours) post bisection of head regeneration and hypostome, tentacle, budding zone, body column, and foot tissues. We identified 27,137 open-chromatin elements in the Hydra genome. To further classify the open-chromatin elements, we carried out ChIP-seq of three histone modifications (H3K4me2, H3K4me3, and H3K27ac) in a subset of the samples corresponding to annotate our ATAC-seq regions as promoter-like or enhancer-like. Integrative analysis of the ATAC-seq and ChIP-seq datasets identified 9998 candidate promoter and 3018 candidate enhancer elements in the Hydra genome. We also show that the distributions of histone marks at these elements resemble those of *cis*-regulatory elements in the bilaterian genomes.

In this study, we provide the first genome-wide atlas and analysis of open-chromatin elements in a cnidarian genome in the context of a developmental process. We report 3018 candidate enhancer-like elements in the Hydra genome, a subset of which are dynamically remodeled during head regeneration. Despite the presence of enhancer-like elements in the

cnidarian genomes, the CTCF gene is missing in the genomes of both *Nematostella vectensis* (Heger, Marin, Bartkuhn, Schierenberg, & Wiehe, 2012) and Hydra (using Blast search of annotated genes). An important future question is to probe the mode of physical interaction of enhancers and promoters in the cnidarian genomes in the absence of CTCF-mediated DNA looping. Additionally, in this study we focused on histone marks associated with active regulatory elements. Future studies on the role of repressive histone marks in gene regulation during important developmental processes, such as head regeneration, can provide further insight.

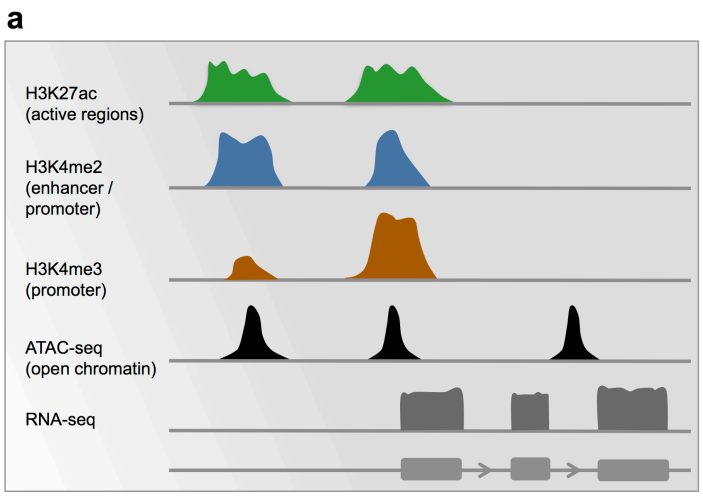
3.5 Figures

Figure 3.1: Schematic of ATAC-seq and ChIP-seq experiments for Hydra head regeneration and various body parts.



For the head regeneration experiment, Hydra were bisected at the boundary of regions R1 and R2 and allowed to regenerate for specific time periods (0, 2, 4, 6, 12, 24, and 48 hours). The region R2 and body parts (foot, body column, budding zone, head, hypostome, tentacles, and whole animal) were harvested for ATAC-seq to map open-chromatin regions or crosslinked with formaldehyde and sonicated for subsequent chromatin immunoprecipitation.

Figure 3.2: Types of high-throughput datasets collected for mapping the *cis*-regulatory modules of Hydra.



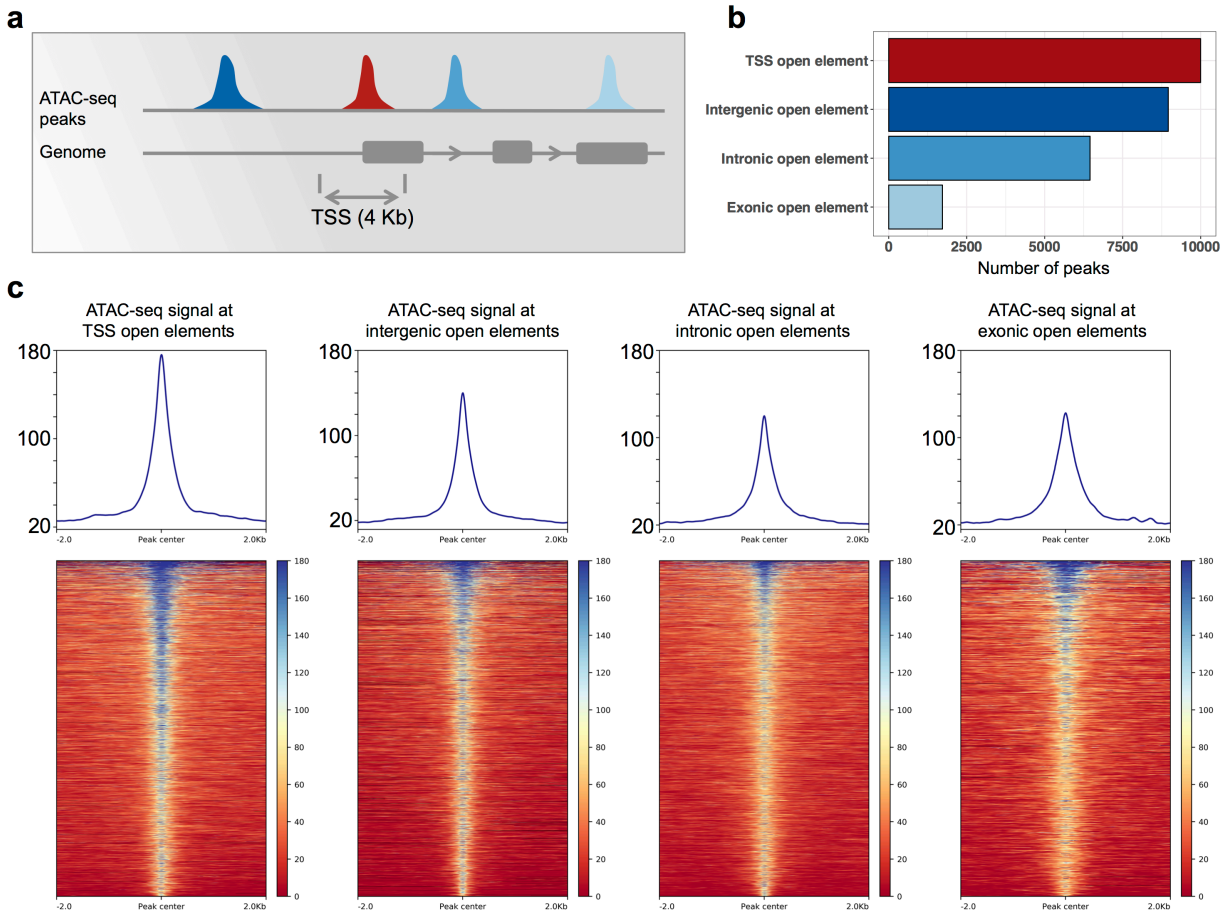
b

Experiment	Samples	RNA	ATAC	ChIP*
Regeneration	0 hr	■	■	■
	2 hr	■	■	
	4hr	■	■	■
	6hr	■	■	■
	12 hr	■	■	
	24 hr	■	■	■
	48 hr	■	■	
Body map	foot	■	■	
	body column	■	■	
	budding zone	■	■	
	hyopstome	■	■	
	tentacle	■	■	
	whole animal			■
	head			■

*ChIP-seq: H3K4me2, H3K4me3, H3K27ac

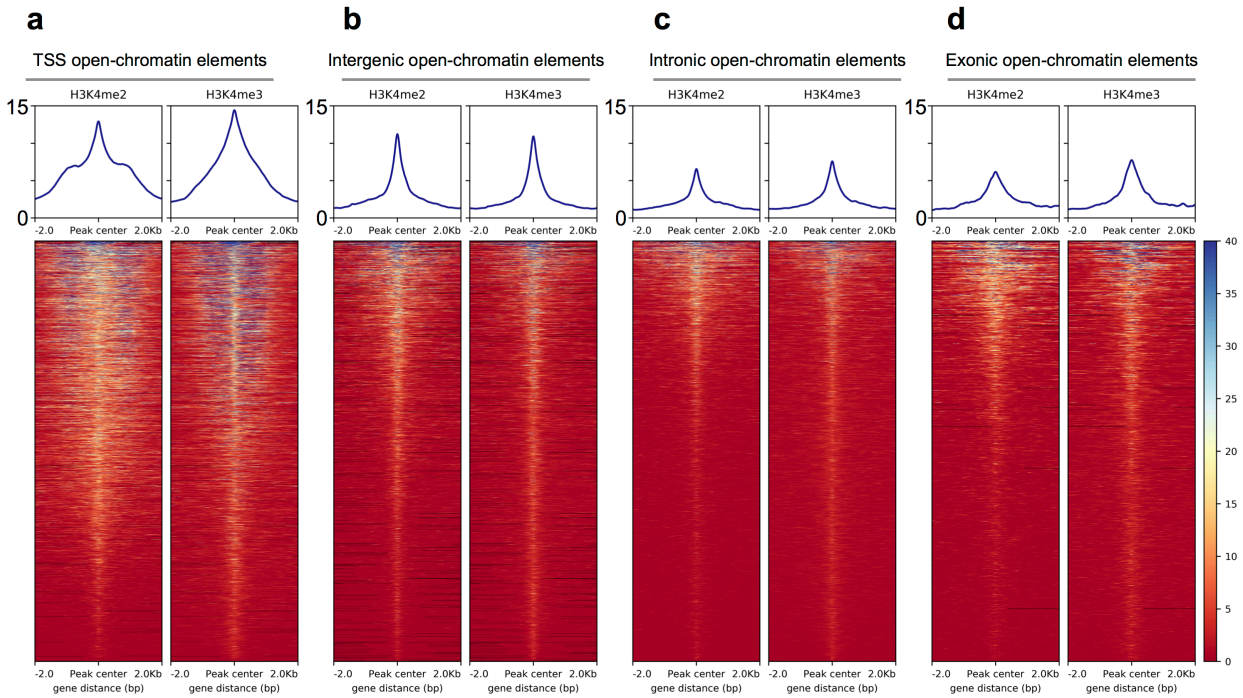
(a) Schematic of histone marks and chromatin accessibility of enhancers and promoters. (b) List of head regeneration and body map samples assayed to measure gene expression (RNA-seq), map open-chromatin regions (ATAC-seq), and map active promoter and enhancer regions using ChIP-seq. Gray boxes indicate samples that were assayed using each of the three techniques.

Figure 3.3: ATAC-seq signals are on average strongest at near TSS and intergenic regions



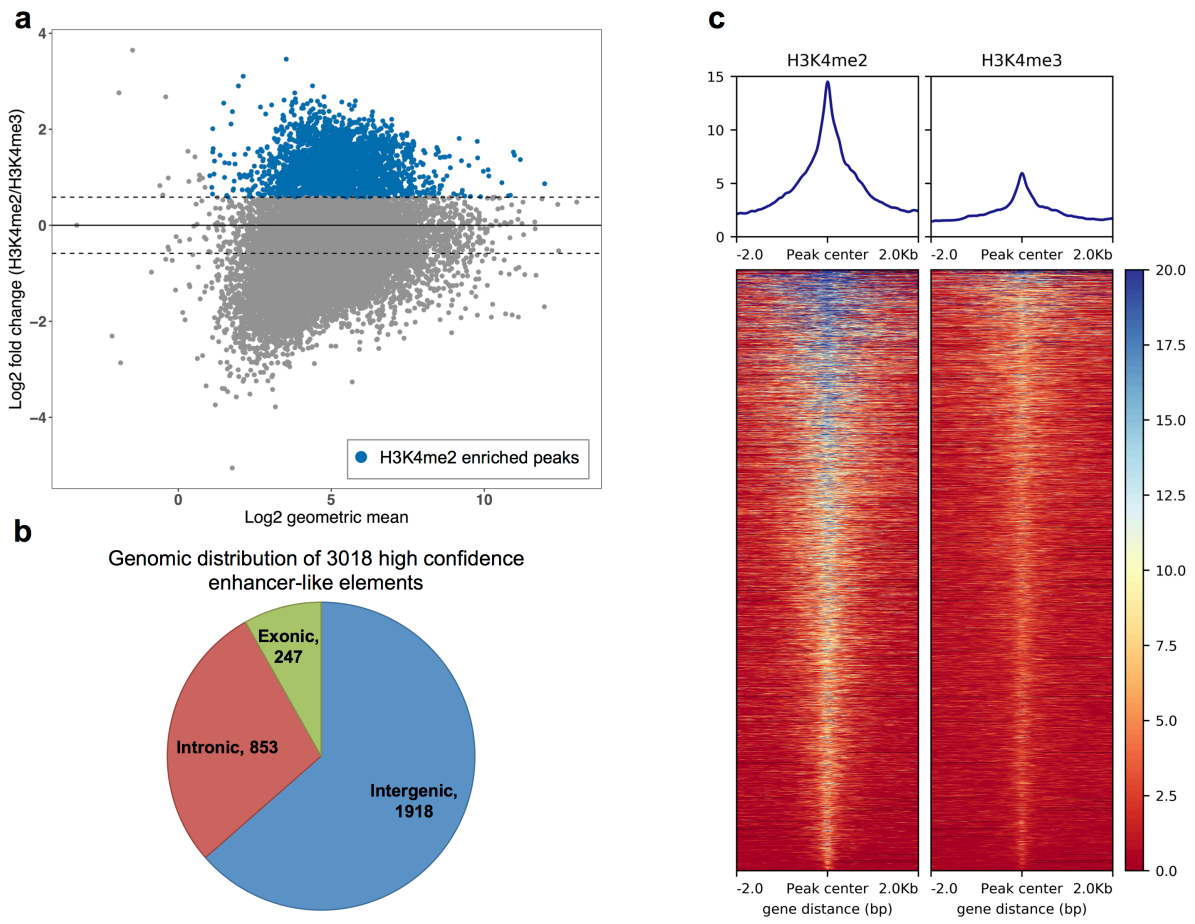
(a) Classification scheme of open-chromatin elements based on genomic location. (b) The 27,137 replicated ATAC-seq peaks called on data from all samples were classified based on overlap with gene loci. The promoter regions were defined as regions within 2 Kb of start of transcripts. (c) ATAC-seq signal enrichment at each type of open-chromatin element. All graphs and heatmaps are to the same scale.

Figure 3.4: Enrichment of histone modification signals at the open-chromatin elements are on average strongest near the TSS and in intergenic regions.



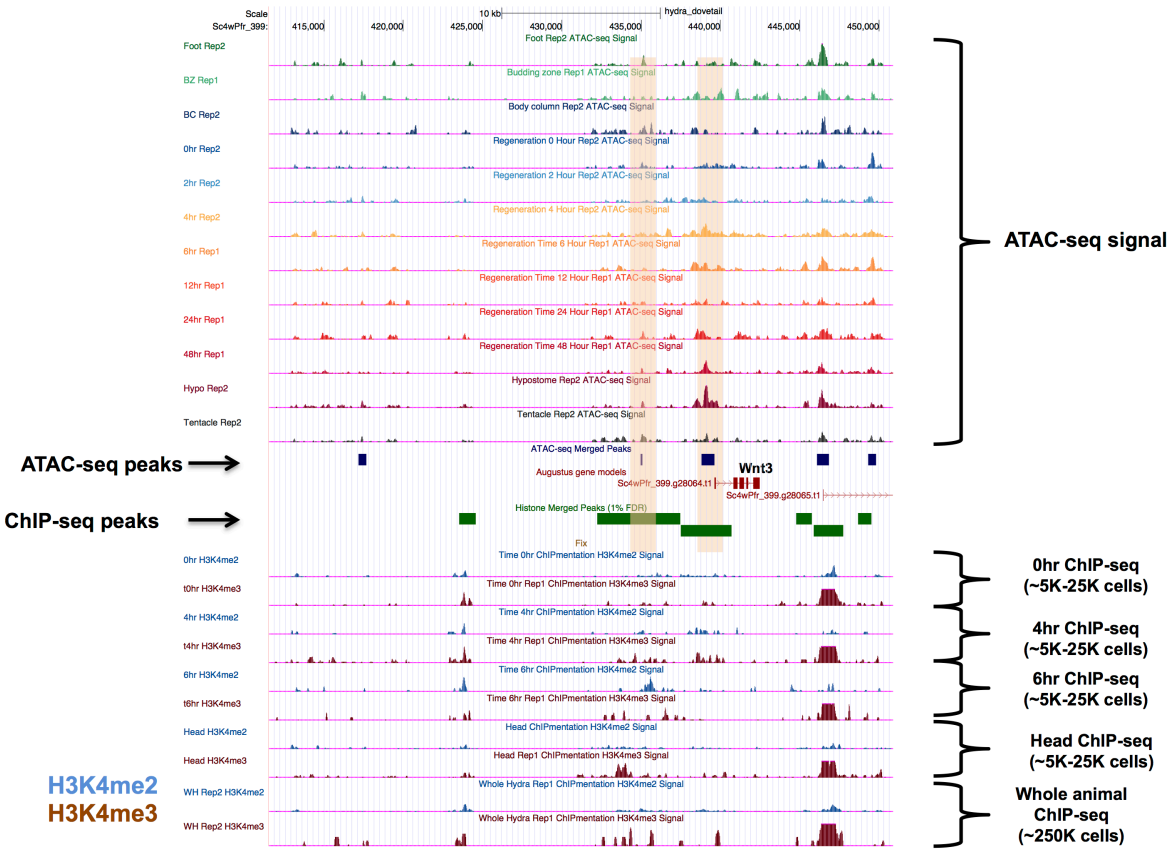
Enrichment of H3K4me2 (first graph and heatmap) and H3K4me3 (second graph and heatmap) histone mark normalized signals at each type of open-chromatin elements: TSS open elements that likely correspond to promoters (a), intergenic open elements (b), intronic open elements (c), and exonic open elements (d). All graphs and heatmaps are to the same scale.

Figure 3.5: Determining a set of candidate enhancer-like elements in Hydra genome using the ratio of H3K4me2 to H3K4me3.



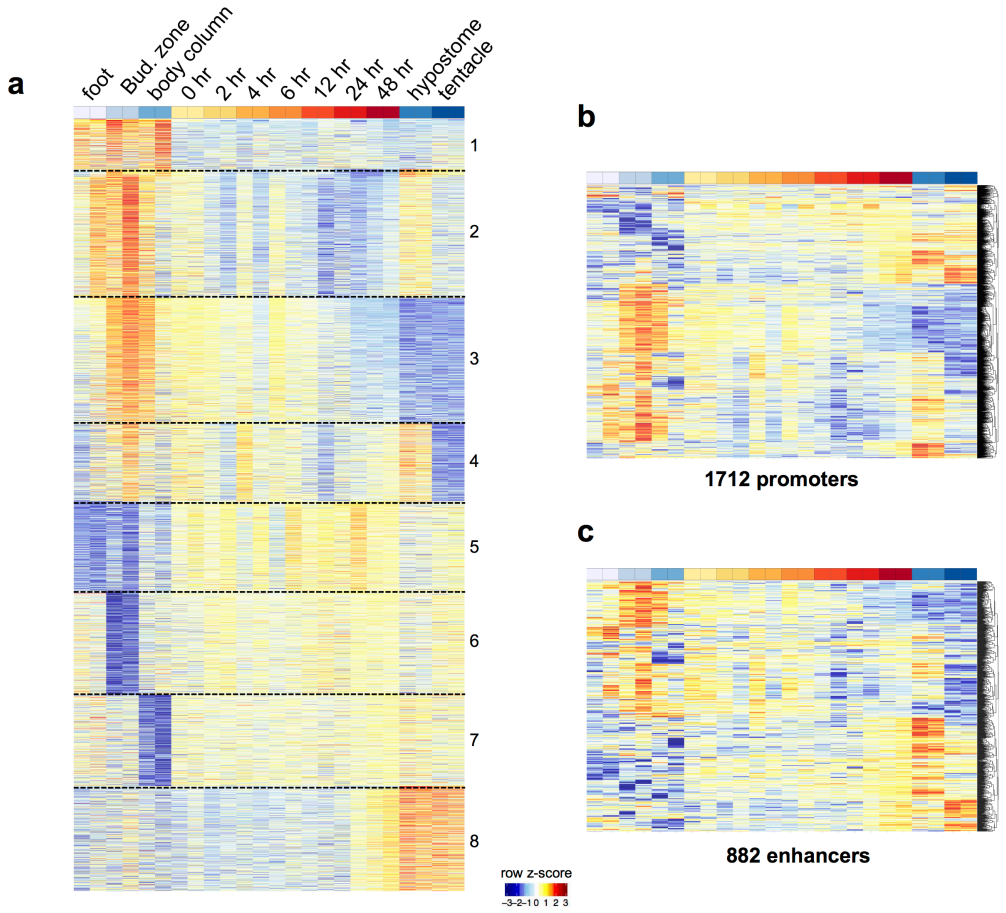
(a) 3018 open-chromatin elements (blue) were enriched in H3K4me2 compared to H3K4me3 to form a set of high-confidence candidate enhancer-like elements. The top dotted line represents 50% higher H3K4me2 than H3K4me3 signal. (b) Genomic distribution of 3018 candidate enhancer-like elements. (c) Enrichment of H3K4me2 and H3K4me3 normalized signals at the 3018 candidate enhancer-like elements.

Figure 3.6: Open-chromatin and ChIP signal tracks for the Wnt3 locus.



Genome browser signal tracks for ATAC-seq and ChIP-seq data are shown for the Wnt3 locus. The Wnt3 promoter and upstream "enhancer" region gain hypersensitivity early in the head regeneration time course.

Figure 3.7: Dynamics of 4168 differentially hypersensitive peaks.



(a) Normalized reads per million (RPM) values for the 4,168 DHS peaks were converted to row z-scores and k-means clustered into 8 clusters based on the observed number of clusters when hierarchically clustered. (b) Hierarchical clustering of 1,712 DHS candidate promoter elements. (c) Hierarchical clustering of 882 DHS candidate enhancer elements.

2.6 Methods and Materials

Hydra culture

Hydra vulgaris polyps were used for the isolation of RNA. They were fed freshly hatched *Artemia salina* nauplii twice per week and cultured as described previously (Smith, Gee, BlitzIl, & Bode, 1999). Animals were starved for at least 1 day before any tissue manipulation, nuclei isolation, or crosslinking.

Experimental design and tissue manipulation

For chromatin profiling experiments using ATAC-seq, the Hydra polyps were first incubated in a cocktail of four antibiotics for one week with feeding, followed by one week of recovery in sterile medium according to the protocol by Fraune *et al.* (Fraune et al., 2015). This was done to remove commensal bacteria from the Hydra polyps before Tn5 tagmentation during ATAC-seq and avoid contamination from tagmented bacterial DNA.

For each sample, 1-day starved asexual hydra polyps were selected. For regeneration, twenty animals per sample (with two biological replicates) were bisected at the region 1 (R1) and region 2 (R2) border (Fig. 3.1) and allowed to undergo head regeneration for a specific period of time (0, 2, 4, 6, 12, 24, or 48 hours). Then the R2 regions of the animals of a sample were isolated for nuclei isolation and tagmentation (ATAC-seq) or crosslinking and immunoprecipitation. The following tissues were collected for the body map samples: ten whole animals; foot, budding zone, body column, hypostome, tentacles, and head (hypostome + tentacles) tissues from 20 animals; regenerating R2 tissues from 20 animals were collected for ATAC-seq and ChIP-seq each.

Chromatin profiling using ATAC-seq

Nuclei from tissues described in the previous section were isolated using the following protocol based on Endl et al. (Endl, Lohmann, Bosch, & Gerhart, 1999)**.** Briefly, Hydra tissues were washed in ice-cold PBS once. The tissues were homogenized in 1 mL of dissociation medium (3.6 mM KCl, 6 mM $CaCl_2$, 1.2 mM $MgSO_4$, 6 mM sodium citrate, 6 mM sodium pyruvate, 6 mM glucose, 12.5 mM TES, stored in 4°C) in a tissue homogenizer. The solution of homogenized tissue was transferred to an eppendorf tube and centrifuged at 500g for 5 min. The supernatant was removed. The cells were resuspended in 50 μL of cold cell lysis buffer (10 mM Tris-HCl, pH 7.4, 10 mM NaCl, 3 mM $MgCl_2$, stored in 4°C) + 0.2% IGEPAL. The sample was immediately spun down at 90g for 8 min. The supernatant was transferred to a fresh eppendorf tube and spun down at 500g for 12 min. The supernatant was removed and the nuclei pellet resuspended in 50 μL ice-cold PBS to remove mitochondrial DNA. The sample was spun down at 500g for 12 min to collect the nuclei for tagmentation. The chromatin in the nuclei pellet was tagmented using 1 μL of Tn5 enzyme and sequencing libraries prepared according to the protocol in Buenrostro et al. (Buenrostro et al., 2013) with the following modification: DNA fragments in the final sequencing library were size selected for 100-500 bp on a 2% agarose gel. The library qualities were assessed using a Bioanalyzer and the libraries were sequenced as 43-bp paired-end reads on an Illumina NextSeq500. Each sample was performed with two biological replicates.

Immunoprecipitation followed by sequencing using ChIP-seq

ChIP-seq libraries were prepared using the ChIPmentation protocol v.1.14 of Schmidl et al. (Schmidl et al., 2015). Briefly, tissues described in the section “Experimental design and

tissue manipulation" were crosslinked in 1% formaldehyde. The crosslinked chromatin samples were sonicated using A Qsonica sonicator using the following settings: 50%, total 15 minutes, 30 seconds on and 30 seconds off. The chromatin samples were sonicated to an average size range of 200-700 bp. The following antibodies were used for immunoprecipitation: H3K4me3 Rabbit mAb (Cell Signaling Technology, Cat. No. 9751), H3K4me2 Rabbit pAb (Millipore-Sigma, Cat. No. 07030), and H3K27ac Rabbit pAb (Active Motif, Cat. No. 39133). Dynabeads M-280 sheep anti-rabbit IgG beads (Thermo Fisher Scientific, Cat. No. 11203D) were used for immunoprecipitation. The sequencing libraries were prepared according to protocol. The library qualities were on a Bioanalyzer and the libraries were sequenced as 43-bp paired-end reads on an Illumina NextSeq500. Each experiment was performed with two biological replicates.

ATAC-seq data analysis

Adapter sequences and low quality base pairs from the paired-ends reads were trimmed using Trimmomatic v. 0.35 (Bolger, Lohse, & Usadel, 2014) using the following parameters: "PE [read1.fastq] [read2.fastq] pe_read1.fastq.gz se_read1.fastq.gz pe_read2.fastq.gz se_read2.fastq.gz ILLUMINACLIP:NexteraPE-PE.fa:2:30:8:4:true LEADING:20 TRAILING:20 SLIDINGWINDOW:4:17 MINLEN:30". The trimmed reads were first mapped to Hydra mitochondrial DNA sequences to filter the mitochondrial reads. The unmapped reads were mapped to the genome sequence from Hydra 2.0 Genome Project (https://research.nhgri.nih.gov/hydra/) using Bowtie v1.2 (Langmead, Trapnell, Pop, & Salzberg, 2009) with the following parameters: -X 2000 -v 3 -m 3 -k 1 –best. Peaks were called using Homer (Heinz et al., 2010) using the following parameters: -size 500 -minDist 50 -fdr 0.01 -style factor. Overlapping peaks between the two replicates of each sample were kept for downstream

analysis. The peaks from all samples were merged using Bedtools v.2.23.0 (Quinlan & Hall, 2010). The read coverage of peaks for each sample were obtained using "coverageBed" function of Bedtools. The read counts for each sample were normalized for efficiency (number of reads within peaks divided by total number of mapped reads) and reads per million. Differentially hypersensitive (DHS) peaks were determined using edgeR's GLM function (Robinson, McCarthy, & Smyth, 2010) at the significance level of 5% FDR and minimum 2-fold change. Signal densities at the peaks were plotted using Deeptools (Ramírez, Dündar, Diehl, Grüning, & Manke, 2014).

ChIP-seq data analysis

Adapter sequences and low quality base pairs from the paired-ends reads were trimmed using Trimmomatic v. 0.35 (Bolger et al., 2014) using the parameters in the previous section. The trimmed reads were mapped to the genome sequence from Hydra 2.0 Genome Project (https://research.nhgri.nih.gov/hydra/) using Bowtie v1.2 (Langmead et al., 2009) with the following parameters: -X 2000 -v 3 -m 3 -k 1 –best. The read coverage of ATAC-seq peaks for each sample were obtained using "coverageBed" function of Bedtools. The read counts for each sample were normalized for efficiency (number of reads within peaks divided by total number of mapped reads) and reads per million. Signal densities of the histone marks at the ATAC-seq peaks were plotted using Deeptools (Ramírez et al., 2014).

3.7 References

Arvizu, F., Aguilera, A., & Salgado, L. M. (2006). Activities of the protein kinases STK, PI3K, MEK, and ERK are required for the development of the head organizer in Hydra magnipapillata. *Differentiation*, *74*(6), 305–312. https://doi.org/10.1111/j.1432-0436.2006.00078.x

Bode, H. R. (2012). The head organizer in Hydra. *International Journal of Developmental Biology*, *56*(6–8), 473–478. https://doi.org/10.1387/ijdb.113448hb

Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics*, *30*(15), 2114–2120. https://doi.org/10.1093/bioinformatics/btu170

Boyle, A. P., Song, L., Lee, B. K., London, D., Keefe, D., Birney, E., … Furey, T. S. (2011). High-resolution genome-wide in vivo footprinting of diverse transcription factors in human cells. *Genome Research*, *21*(3), 456–464. https://doi.org/10.1101/gr.112656.110

Buenrostro, J. D., Giresi, P. G., Zaba, L. C., Chang, H. Y., & Greenleaf, W. J. (2013). Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nature Methods*, *10*(12), 1213–1218. https://doi.org/10.1038/nmeth.2688

Bulger, M., & Groudine, M. (1999). Looping versus linking: Toward a model for long-distance gene activation. *Genes and Development*, *13*(19), 2465–2477. https://doi.org/10.1101/gad.13.19.2465

Carroll, S. B. (2008). Evo-Devo and an Expanding Evolutionary Synthesis: A Genetic Theory of Morphological Evolution. *Cell*, *134*(1), 25–36. https://doi.org/10.1016/j.cell.2008.06.030

Chapman, J. A., Kirkness, E. F., Simakov, O., Hampson, S. E., Mitros, T., Weinmaier, T., … Steele, R. E. (2010). The dynamic genome of Hydra. *Nature*, *464*(7288), 592–596. https://doi.org/10.1038/nature08830

Consortium, E. P., Dunham, I., Kundaje, A., Aldred, S. F., Collins, P. J., Davis, C. a, … Lochovsky, L. (2012). An integrated encyclopedia of DNA elements in the human genome. *Nature*, *489*(7414), 57–74. https://doi.org/10.1038/nature11247

Endl, I., Lohmann, J. U., Bosch, T. C. †G. G., & Gerhart, J. C. (1999). Head-specific gene expression in Hydra: Complexity of DNA- protein interactions at the promoter of ks1 is inversely correlated to the head activation potential. *Pnas*, *96*(4), 1445–1450. https://doi.org/10.1073/pnas.96.4.1445

Frankel, N., Erezyilmaz, D. F., McGregor, A. P., Wang, S., Payre, F., & Stern, D. L. (2011). Morphological evolution caused by many subtle-effect substitutions in regulatory DNA. *Nature*, *474*(7353), 598–603. https://doi.org/10.1038/nature10200

Fraune, S., Anton-Erxleben, F., Augustin, R., Franzenburg, S., Knop, M., Schröder, K., … Bosch, T. C. G. (2015). Bacteria-bacteria interactions within the microbiota of the ancestral metazoan Hydra contribute to fungal resistance. *ISME Journal*, *9*(7), 1543–1556. https://doi.org/10.1038/ismej.2014.239

Heger, P., Marin, B., Bartkuhn, M., Schierenberg, E., & Wiehe, T. (2012). The chromatin insulator CTCF and the emergence of metazoan diversity. *Proceedings of the National Academy of Sciences*, *109*(43), 17507–17512. https://doi.org/10.1073/pnas.1111941109

Heinz, S., Benner, C., Spann, N., Bertolino, E., Lin, Y. C., Laslo, P., … Glass, C. K. (2010). Simple Combinations of Lineage-Determining Transcription Factors Prime cis-Regulatory Elements Required for Macrophage and B Cell Identities. *Molecular Cell*, *38*(4), 576–589.

https://doi.org/10.1016/j.molcel.2010.05.004
Hesselberth, J. R., Chen, X., Zhang, Z., Sabo, P. J., Sandstrom, R., Reynolds, A. P., … Stamatoyannopoulos, J. A. (2009). Global mapping of protein-DNA interactions in vivo by digital genomic footprinting. *Nature Methods*, *6*(4), 283–289. https://doi.org/10.1038/nmeth.1313
Langmead, B., Trapnell, C., Pop, M., & Salzberg, S. L. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biology*, *10*(3). https://doi.org/10.1186/gb-2009-10-3-r25
Lengfeld, T., Watanabe, H., Simakov, O., Lindgens, D., Gee, L., Law, L., … Holstein, T. W. (2009). Multiple Wnts are involved in Hydra organizer formation and regeneration. *Developmental Biology*, *330*(1), 186–199. https://doi.org/10.1016/j.ydbio.2009.02.004
Lettice, L. A., Heaney, S. J. H., Purdie, L. A., Li, L., de Beer, P., Oostra, B. A., … de Graaff, E. (2003). A long-range Shh enhancer regulates expression in the developing limb and fin and is associated with preaxial polydactyly. *Human Molecular Genetics*, *12*(14), 1725–1735. https://doi.org/10.1093/hmg/ddg180
Levine, M. (2010). Transcriptional enhancers in animal development and evolution. *Current Biology*, *20*(17), R754–R763. https://doi.org/10.1016/j.cub.2010.06.070
Maston, G. A., Evans, S. K., & Green, M. R. (2006). Transcriptional Regulatory Elements in the Human Genome. *Annual Review of Genomics and Human Genetics*, *7*(1), 29–59. https://doi.org/10.1146/annurev.genom.7.080505.115623
Neph, S., Vierstra, J., Stergachis, A. B., Reynolds, A. P., Haugen, E., Vernot, B., … Stamatoyannopoulos, J. A. (2012). An expansive human regulatory lexicon encoded in transcription factor footprints. *Nature*, *489*(7414), 83–90. https://doi.org/10.1038/nature11212
Petersen, H. O., Hï¿½ger, S. K., Looso, M., Lengfeld, T., Kuhn, A., Warnken, U., … Holstein, T. W. (2015). A comprehensive transcriptomic and proteomic analysis of hydra head regeneration. *Molecular Biology and Evolution*, *32*(8), 1928–1947. https://doi.org/10.1093/molbev/msv079
Putnam, N. H., Srivastava, M., Hellsten, U., Dirks, B., Chapman, J., Salamov, A., … Rokhsar, D. S. (2007). Sea Anemone Genome Reveals Ancestral Eumetazoan Gene Repertoire and Genomic Organization. *Science*, *317*(5834), 86–94. https://doi.org/10.1126/science.1139158
Quinlan, A. R., & Hall, I. M. (2010). BEDTools: A flexible suite of utilities for comparing genomic features. *Bioinformatics*, *26*(6), 841–842. https://doi.org/10.1093/bioinformatics/btq033
Ramírez, F., Dündar, F., Diehl, S., Grüning, B. A., & Manke, T. (2014). DeepTools: A flexible platform for exploring deep-sequencing data. *Nucleic Acids Research*, *42*(W1). https://doi.org/10.1093/nar/gku365
Robinson, M. D., McCarthy, D. J., & Smyth, G. K. (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics (Oxford, England)*, *26*(1), 139–40. https://doi.org/10.1093/bioinformatics/btp616
Roy, S., Ernst, J., Kharchenko, P. V., Kheradpour, P., Negre, N., Eaton, M. L., … Lowdon, R. F. (2010). Identification of functional elements and regulatory circuits by Drosophila modENCODE. *Science*, *330*(6012), 1787–1797. https://doi.org/10.1126/science.1198374
Schmidl, C., Rendeiro, A. F., Sheffield, N. C., & Bock, C. (2015). ChIPmentation: Fast, robust, low-input ChIP-seq for histones and transcription factors. *Nature Methods*, *12*(10), 963–965. https://doi.org/10.1038/nmeth.3542

Schwaiger, M. (2014). Evolutionary conservation of the eumetazoam gene regulatory landscape - Supplemental Figures. *Genome Research*, 1–13. https://doi.org/10.1101/gr.162529.113.Freely

Smith, K. M., Gee, L., BlitzIl, & Bode, H. R. (1999). CnOtx, a mamber of the OTX gene family, has a role in cell movement in Hydra. *Developmental Biology*, *212*, 392–404.

Technau, U., & Steele, R. E. (2012). Evolutionary crossroads in developmental biology: Cnidaria. *Development*, *139*(23), 4491–4491. https://doi.org/10.1242/dev.090472

Villar, D., Berthelot, C., Aldridge, S., Rayner, T. F., Lukk, M., Pignatelli, M., … Odom, D. T. (2015). Enhancer evolution across 20 mammalian species. *Cell*, *160*(3), 554–566. https://doi.org/10.1016/j.cell.2015.01.006

Visel, A., Taher, L., Girgis, H., May, D., Golonzhka, O., Hoch, R. V., … Rubenstein, J. L. R. (2013). A high-resolution enhancer atlas of the developing telencephalon. *Cell*, *152*(4), 895–908. https://doi.org/10.1016/j.cell.2012.12.041

Chapter 4

Comparative dynamics of miRNA expression during mouse and human prenatal development

Statement of contribution

In this study, mouse tissues were harvested and RNA extracted at California Institute of Technology by Brian Williams (Wold Lab). Mouse miRNA-seq libraries were prepared and sequenced at HudsonAlpha Institute for Biotechnology (Myers Lab). Mouse RNA-seq libraries were prepared and sequenced at California Institute of Technology (Wold Lab). Human short RNA-seq samples were prepared at Cold Spring Harbor Laboratory (Gingeras Lab). Mouse NanoString samples were prepared by Dr. Weihua Zeng at University of California Irvine (Mortazavi Lab). Mouse ChIP-seq samples were prepared at Univeristy of California San Diego (Ren Lab). I analyzed and interpreted results for mouse miRNA data. Alessandra Breschi (CRG Spain, currently at Stanford University) analyzed and interpreted the human miRNA data and performed the human-mouse comparative study.

4.1 Abstract

microRNAs (miRNAs) are a class of small non-coding RNA that are critical post-transcriptional regulators of gene expression. The ENCODE project profiled the expression of miRNAs in various tissues during a time-course of embryonic development in mouse and several human prenatal samples using multiple complementary sequencing and hybridization techniques. We detected the expression of 671 miRNAs in mouse as well as 388 miRNAs in human. We find distinct tissue and developmental stage specific miRNA expression profiles dominated by a small number of miRNAs. Comparative analysis of conserved miRNAs reveals clustering of expression patterns by tissue types rather than species. We used matching messenger RNA-seq (mRNA-seq) and histone modification ChIP-seq datasets to improve the annotation of miRNA primary transcripts. We show that the expression levels of a majority of primary miRNA transcripts predict the expression of their corresponding mature miRNAs. Our data provide the most comprehensive miRNA resource for mouse and human embryonic development as well as a comprehensive list of mouse miRNAs that can be reliably measured by mRNA-seq of their primary transcripts.

4.2 Introduction

microRNAs (miRNAs) are small ~22 nucleotide (nt) endogenous non-coding RNAs that regulate gene expression by mediating the post-transcriptional degradation of messenger RNA (mRNA) or hindering the translation of proteins (Bartel, 2004; He & Hannon, 2004). miRNA biogenesis occurs in several steps. Typically, polyadenylated primary miRNA (pri-miRNA) transcripts (>200 nt), which are sometimes referred to as “host genes” and have a characteristic hairpin structure, are transcribed by RNA polymerase II. The pri-miRNA are cleaved in the

nucleus by the enzyme Drosha into pre-miRNA (~80 nt) that are exported to the cytoplasm and finally processed into 18-28 nt mature miRNA by the enzyme Dicer (Han et al., 2006). The first miRNA was discovered in the nematode *C. elegans* in 1993 (Lee, Feinbaum, & Ambros, 1993) but during over two decades since, thousands of miRNAs have been discovered and annotated in diverse plants, metazoans, and some viruses (Kozomara & Griffiths-Jones, 2011). Studies have shown that most genes are potential targets of miRNAs (Friedman, Farh, Burge, & Bartel, 2009) and that miRNAs are involved in regulating diverse cellular processes during development and homeostasis (Vidigal & Ventura, 2015). Dysregulation of miRNA expression is known to underlie numerous diseases and developmental defects such as cancer (Lin & Gregory, 2015), cardiovascular diseases (Romaine, Tomaszewski, Condorelli, & Samani, 2015; Zhao et al., 2015), and neurological diseases (Cao, Li, & Chan, 2016).

miRNAs have been profiled in various tissues and primary cells in diverse metazoans and plants (Ehrenreich & Purugganan, 2008; Lagos-Quintana et al., 2002; Wienholds et al., 2005) . Although previous studies have described the tissue-specificity of a large number of miRNAs in multiple organisms (Guo et al., 2014; Lagos-Quintana et al., 2002; Ludwig et al., 2016), a characterization of miRNA temporal dynamics across mammalian development is still mostly missing. The ENCODE Consortium is using DNA and RNA sequencing technologies to catalogue the functional elements in the human (Encode Consortium, 2012) and mouse (Yue et al., 2014) genomes. Previous efforts by the Consortium focused on a meta-analysis of 501 human and 236 mouse small RNA sequencing data sets from a multitude of published sources leading to the characterization of splicing-derived miRNAs (mirtrons) in the human and mouse genomes (Ladewig, Okamura, Flynt, Westholm, & Lai, 2012). However, the diversity of the

source tissues and the different underlying experimental protocols from the various underlying primary sources complicated any sort of systematic quantitative analysis.

With growing evidence of the critical role of miRNAs in homeostasis and disease, multiple techniques have been developed for profiling the expression of mature miRNAs, each with their own strengths (Mestdagh et al., 2014). RNA-seq typically refers to the profiling of expressed transcripts 200 nt or longer including the messenger RNAs (mRNA) and long non-coding RNAs (lncRNA) (Mortazavi, Williams, McCue, Schaeffer, & Wold, 2008), which in this work we will refer to as messenger RNA-seq (mRNA-seq), whereas there are also multiple miRNA-specific sequencing protocols such as microRNA-seq (Roberts et al., 2015) and short RNA-seq (Fejes-Toth et al., 2009). There are also hybridization-based assays such as microarrays as well as molecule counting such as NanoString, which involves hybridization of color-coded molecular barcodes (Geiss et al., 2008; Wyman et al., 2011). As mature miRNAs are processed from longer host pri-miRNAs and the annotated pri-miRNAs are predominantly protein-coding or lncRNA transcripts (Cai, Hagedorn, & Cullen, 2004), we hypothesize that mRNA-seq should be able to profile the expression of pri-miRNAs. However, there is a significant number of miRNAs whose host genes have not been characterized yet. Furthermore, an important question is whether the expression of pri-miRNAs can reliably predict the expression of their corresponding mature miRNAs. This would allow the simultaneous profiling of mature miRNA expression along with mRNAs using mRNA-seq. Previous studies have attempted to answer this question in specific cell types (Zeng et al., 2016). Availability of matching mRNA-seq and microRNA-seq data sets for the same samples in our study provides a unique opportunity to answer this question.

In this study, we have used microRNA-seq and NanoString for 16 different mouse tissues at 8 embryonic (e10.5 – P0) stages and short RNA-seq for 17 human tissues in 7 stages (week 19-40) of fetal development. Integrative analysis of these data sets with matching ENCODE data allowed us to compare the characteristics and dynamics of miRNA expression during mouse and human development. In particular, comparisons of matching mRNA-seq and microRNA-seq data sets allowed us to identify novel pri-miRNA transcripts and to answer the question of how accurately the expression of pri-miRNAs predict the expression of their corresponding mature miRNAs.

4.3 Results

A reference miRNA catalog across mammalian development

As part of the ENCODE project, we used microRNA-seq and NanoString to profile mature miRNAs in mouse, short RNA-seq to profile pre-miRNAs and mature miRNAs in human, and mRNA-seq to profile the expression of pri-miRNAs across different stages of human and mouse development (Fig. 4.1a). This study encompasses 156 microRNA-seq and 154 NanoString datasets in mouse, 32 short RNA-seq datasets in human, and 156 mRNA-seq datasets in mouse and human (Fig. 4.1b). In addition, 2 human cells lines (GM12878 and K562) were assayed using all technologies for comparison. All the data supporting this study are available from the ENCODE data portal (www.encodeproject.org).

We used the microRNA-seq and short RNA-seq reads to identify novel miRNAs in mouse and human. We pooled the samples by tissue type and discovered novel miRNAs, which we required to be discovered independently in at least two tissues and detected in at least one sample at a minimum of two counts per million in order to be retained for downstream analysis

[Methods]. This analysis gave us sets of 72 and 83 novel miRNAs in mouse and human respectively.

We used multiple techniques to profile miRNAs in mouse and human and used the data sets to perform a comparative analysis of miRNA expression across development in matching samples. We compared the different techniques in human K562 and GM12878 cells and compared them to datasets from a previous phase of ENCODE. We show that there is high correlation between microRNA-seq and short RNA-seq (Pearson correlation = 0.81) (Fig. 4.1c) and microRNA-seq and NanoString (median Spearman correlation = 0.68) (Fig. 4.7), which matches reproducibility between platforms (Mestdagh et al., 2014). This level of reproducibility clearly allows us to differentiate between different cell types, even across different methods and batches (Fig. 4.1d).

miRNA landscape across mouse embryonic development

We used microRNA-seq reads to quantify miRNA expression levels in mouse using GENCODE annotations version M10, which includes 2202 miRNAs as well as our 72 novel miRNAs. Global analysis of mouse tissues and developmental stages shows distinct miRNA expression patterns as revealed by principal component analysis (PCA) (Fig. 4.2a). Principal component 1 (PC1) clearly separates the various tissues with the nervous system and liver tissues at the extremes, whereas PC2 captures the time component of mouse developmental time-course with a temporal gradient between early development at embryonic day 10 (e10.5) and postnatal samples right after birth (P0) (Fig. 4.2a). There are no significant differences in the number of distinct miRNAs expressed in mouse tissues and developmental stages although the P0 stage shows the least number of distinct miRNAs even though it has the highest number of tissues

profiled (Fig. 4.2b). This result is in contrast to the finding that the absolute numbers of expressed miRNAs increase over developmental time in other model organisms such as *Drosophila melanogaster* (Ninova, Ronshaugen, Griffiths-jones, & Griffiths-jones, 2014). At the tissue level, we find that the nervous system and heart samples show the highest number of distinct miRNAs expressed (Fig 4.2c).

As expected, miRNA expression in mouse tissues is dominated by the expression of a few highly expressed miRNAs with the top 100 miRNAs, ranked by expression, accounting for more than 90 percent of reads in every tissue (Fig 4.2d). But unlike adult mouse tissues in which the corresponding tissue-specific miRNAs are the top expressed miRNAs (Lagos-Quintana et al., 2002), the majority of top expressed miRNAs in the developing tissues are usually ubiquitous and non-tissue specific miRNAs such as Mir-335 and Mir-99a. Liver and kidney are notable exceptions to this trend in which Mir-122 and Mir-10a are the highest expressed miRNAs respectively. The top 10 ranked miRNAs in each tissue contain well-known examples of corresponding tissue-specific miRNAs such as Mir-122 in liver, Mir-9 in brain samples, Mir-10a and Mir-10b in kidney. However, we do not detect high levels of other well-characterized miRNAs such as Mir-1 in muscle, which are more highly expressed after birth.

miRNA expression during mouse embryonic and human fetal development

Previous studies have described multiple tissue-specific miRNAs for major tissue types in a wide variety of model organisms (Guo et al., 2014). Although there are well-described examples of tissue-specific miRNAs, a characterization of developmental stage specific miRNAs is lacking. We performed a time-series analysis of our time-course data for mouse embryonic development to determine the tissue-specific as well as stage-specific miRNAs.

We obtained 23 clusters of tissue-specific and stage-specific miRNAs (Fig. 4.3 and Fig. 4.8) consisting of 363 GENCODE and 9 novel miRNAs that are differentially expressed. 13 of these clusters consist of tissue-specific miRNAs, of which 3 clusters (clusters 13, 17, and 22) are specific to only one tissue while 10 clusters are specific to two or more tissues. 16 clusters show stage-specific dynamics. Of these, 5 clusters (cluster 4, 5, 9, 10, and 12) are higher in early developmental time points while 11 clusters (cluster 1, 2, 6, 7, 8, 13, 18, 19, 20, 21, and 22) consist of miRNAs with increasing expression levels across mouse development in various tissues. 7 clusters of miRNAs exhibit complex dynamics across tissues and developmental stages.

We recovered multiple brain-specific miRNA clusters (#1, 2, 3, and 4) with 43, 9, 31, and 14 miRNAs respectively. Clusters 1, 2, and 3 consist of miRNAs that increase in expression along development while the miRNAs in cluster 4 are highly expressed in early stages (e10.5-e12.5) and decrease in expression in later stages. These three clusters include well-known examples of brain specific miRNAs such as Mir-124a, Mir-125b, Mir-137, Mir-138, Mir-132, Mir-7, Mir-128, Mir-129, Mir-153, Mir-212, and Mir-330. We show that of the brain-specific miRNAs, 14 are specific to early developmental stages compared to 83 miRNAs that increase in expression along the developmental stages. Similarly, we show that heart-specific Mir-302a, Mir-302b, Mir-302c, and Mir-302d (cluster 12) are specific to early time points of the mouse development (e10.5 – e12.5), while Mir-1a-2, Mir-208, and Mir-133a-1 increase in expression at later time points (cluster 13). We generally find miRNAs specific to other tissues such as liver, lung, craniofacial prominence, intestine, and stomach that also exhibit dynamic temporal expression patterns across development. Thus, our analysis provides a comprehensive atlas of stage-specific temporal dynamics of miRNAs across mouse embryonic development.

We used NanoString, which includes a panel of probes for 600 of known miRNAs in mouse, to profile mature miRNAs in the same mouse samples as microRNA-seq for validation purposes. A similar time-series analysis of the NanoString data produced 15 clusters of differentially expressed miRNAs. Comparison of the miRNA clusters from NanoString and microRNA-seq reveals that for 10 of the NanoString clusters, 50% or more of miRNAs overlap between the clusters with similar expression profiles (Fig. 4.9 & Fig. 4.10).

Similarly, we quantified human known and novel miRNAs using short RNA-seq and GENCODE v.25 annotations consisting of 1569 known miRNAs supplemented with the novel miRNAs. A global PCA analysis of miRNA expression shows the brain samples clustering as previously seen in mouse (Fig. 4.4a). While the availability of human samples was more limited compared to mouse, we identified 279 tissue-specific miRNAs, most of which (83%) are preferentially expressed in neuronal and muscular tissues (Fig. 4.4b). As in mouse, brain-specific miRNAs include several well-known examples, such as Mir-9, Mir-124 and Mir-125, while Mir-1, Mir-133, Mir-196 and Mir-206 are well-known muscle-specific miRNAs. While Mir-9 and Mir-125 are among the top 10 expressed miRNAs, other miRNAs such as Mir-1 and Mir-124 are more lowly expressed than in the adult similar as in mouse.

Comparative dynamics of conserved miRNAs during development

In order to compare miRNA expression across development, we first searched for orthologous miRNAs between mouse and human (Fig. 4.5a). We found that a subset of miRNAs is conserved between mouse and human, with 304 miRNAs having a one-to-one orthologous relationship. Our analysis also revealed that 838 and 516 miRNAs in human and mouse respectively lack a clear ortholog in the other species.

We detected only a fraction of the miRNAs in each species although we profiled miRNAs in diverse tissues and developmental time points in mouse and human. Approximately 26% and 23% of GENCODE annotated miRNAs were detected in at least one tissue in mouse and human respectively (Fig. 4.5b). We detected 600 known and 72 novel miRNAs in mouse and 366 known and 83 novel miRNAs in human at a minimum of 2 reads per million in at least one sample. On average, 431 and 218 miRNAs were expressed in each mouse and human sample respectively.

We used the set of one-to-one miRNA orthologs to perform PCA on matching tissues in mouse and human, which revealed clustering of samples based on tissue types (Fig 4.5c). Furthermore, comparative clustering of tissues in mouse and human reveals distinct miRNA expression patterns similar to the clustering of tissues in mouse only. As in the mouse developmental time course, the nervous system and liver tissues cluster separately from the rest of the tissues.

We compared the sets of tissue-specific orthologous miRNAs across all the available tissues in mouse and human, and represented each comparison as pie charts, where the sizes of the pie charts are in proportion to the number of tissue-specific miRNAs (Fig 4.5d). We found that the muscle tissues in mouse and human show the highest conservation (>50%) of expression while the conservation of expression among the corresponding brain, neural tube and lung tissues is significant (~50%), whereas the conservation of expression between the liver samples is low (Fig 4.5d).

Correlation of expression among pri-miRNAs and their corresponding mature miRNAs

The availability of matching microRNA-seq and mRNA-seq data allowed us to evaluate whether the expression of pri-miRNAs is predictive of the expression of their corresponding mature miRNAs. Less than 50% of miRNAs in mouse have annotated primary transcripts in GENCODE version M10 (Fig. 4.6b). We used mRNA-seq data to assemble additional transcript models and supplement the GENCODE annotations, which increased the number of pri-miRNAs in mouse and human by 7% and 17% percent respectively. A representative novel model transcript in mouse, assembled using all mouse mRNA-seq data sets (Methods), overlaps Mir-let7a and Mir-let7f that were lacking annotated pri-miRNAs and is supported by stage-matched chromatin immunoprecipitation followed by sequencing (ChIP-seq) for both H3K4me3 marking the putative promoter (Fig. 4.11) and H3K36me3, as correlate of transcription (Fig. 4.6a). Global correlation analysis of the expression levels of the pri-miRNAs (measured by mRNA-seq) and their corresponding miRNAs (measured by microRNA-seq) shows that 143 (41% of the miRNAs expressed at a minimum of 10 CPM) are well correlated (Spearman correlation ≥ 0.6) with their corresponding pri-miRNAs across the developmental time-course. The median Spearman correlation for all the miRNAs and their corresponding pri-miRNAs is 0.51 (Fig. 4.6c,d). Thus, mRNA and miRNA expression are tightly coupled and miRNA expression can be imputed from the expression of its primary transcript.

4.4 Discussion

In this study we provide a comprehensive resource of miRNA expression dynamics across human and mouse development in multiple tissues. Our catalogue of tissue and developmental stage specific miRNAs provides a valuable resource for elucidating the role of miRNAs in shaping the landscape of mammalian development.

Our analysis highlights certain key properties of miRNAs across mammalian development. Although we assayed 16 and 17 different tissues that are representative of major organ systems in mouse and human respectively across multiple developmental stages, we detected only ~25% of the annotated miRNAs in each species. This result suggests that only a small subset of miRNAs might be involved in regulating gene expression during mammalian development with the remaining either expressed in other tissues or more likely expressed later in post-birth development and adult tissues. This is in contrast with other studies, which have detected the majority of miRNAs in adult tissues (Ludwig et al., 2016). There is also slight variability in the number of miRNAs detected per tissues with the heart and nervous system tissues exhibiting the highest number of detected miRNAs. Interestingly, the miRNA output of most embryonic samples is dominated by the expression of a few highly expressed miRNAs that usually consist of non-tissue-specific and ubiquitously but highly expressed miRNAs.

Although tissue specificity of miRNAs has been well studied and well reported in multiple model organisms (Gao et al., 2011; Lagos-Quintana et al., 2002; Ludwig et al., 2016), a comprehensive study of the dynamics of such tissue-specific miRNAs across mammalian development was lacking. Our analysis fills this knowledge gap. We show that most of the tissue-specific miRNAs are dynamically regulated across development, with different subsets of miRNAs in the brain and heart expressed at different levels during embryonic development.

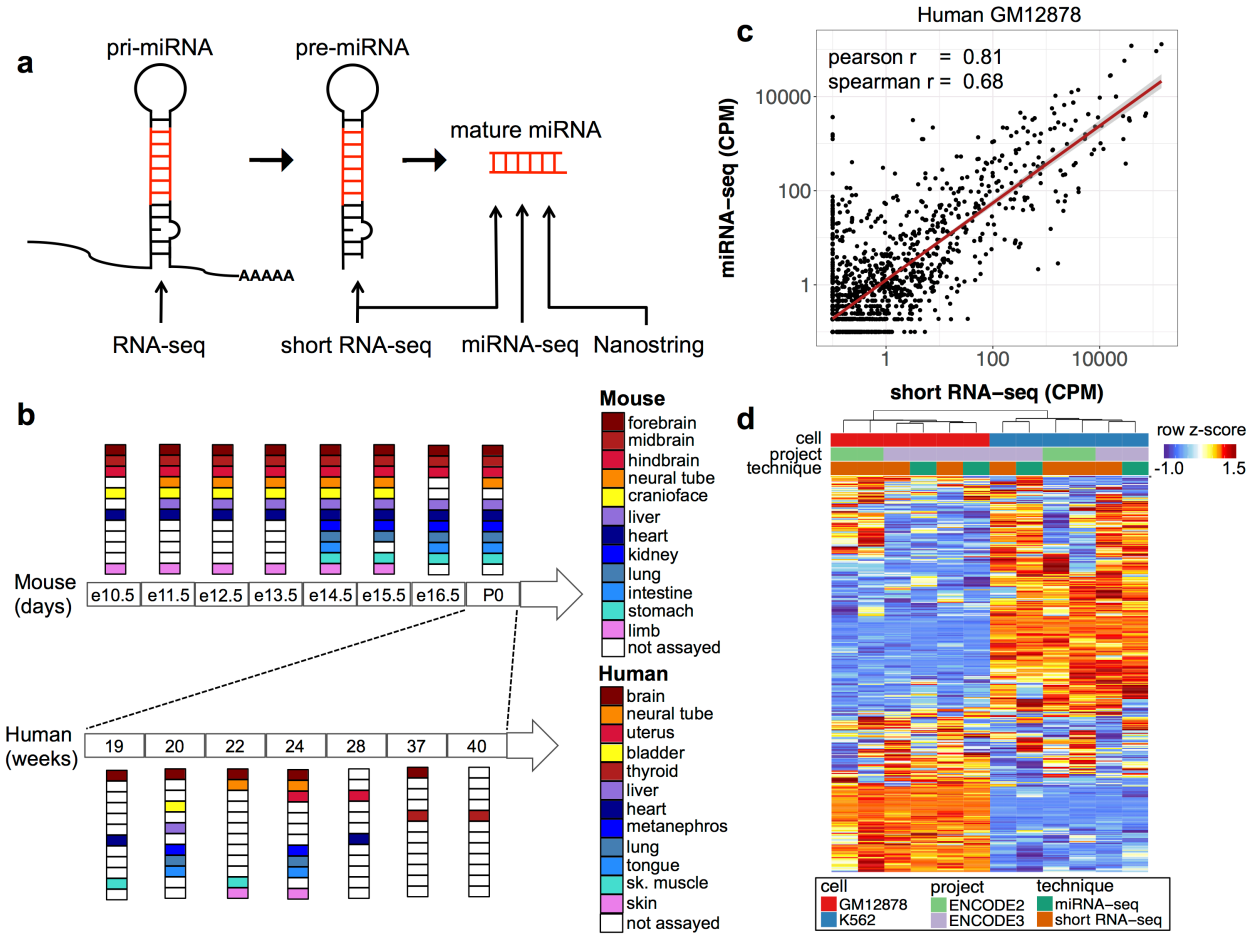
We provide evidence that the tissue-specific expression of a subset of miRNAs is conserved in human and mouse although the overall transcriptional programs are known to have considerably diverged in the two species (Yue et al., 2014). Although the number of one-to-one miRNA orthologs in human and mouse is low as a fraction of the known miRNAs in the two species (~20% of annotated miRNAs in human have one-to-one orthologs in mouse), we show

that the tissue-specific expression patterns of the miRNA orthologs closely resemble the overall patterns observed in each individual species. The conservation of miRNA expression in human and mouse tissues is driven by core sets of tissue-specific miRNAs. We show that the expression of tissue-specific miRNAs is well conserved in some tissues (brain, muscle, and lung) while less conserved in other tissues (liver). The fraction of conserved miRNAs is significantly lower than the number of conserved genes between human and mouse (Herrero et al., 2016), which suggests that miRNAs are evolving more frequently.

mRNA-seq ("regular" RNA-seq) is the most widely adopted assay for profiling the expression of transcripts, while specialized protocols are used to profile the small RNA subset of the transcriptome. While these are normally separate experiments, where the analysis of mRNA-seq is restricted to the long genes (protein-coding or long noncoding RNAs), we show that the expression levels of the primary transcripts of miRNAs as measured by mRNA-seq is highly predictive of the expression of their corresponding mature miRNAs. We should therefore be able to confidently predict the expression of miRNAs when they have reliable host gene models, which opens a much wider set of mRNA-seq samples to miRNA analysis.

4.5 Figures

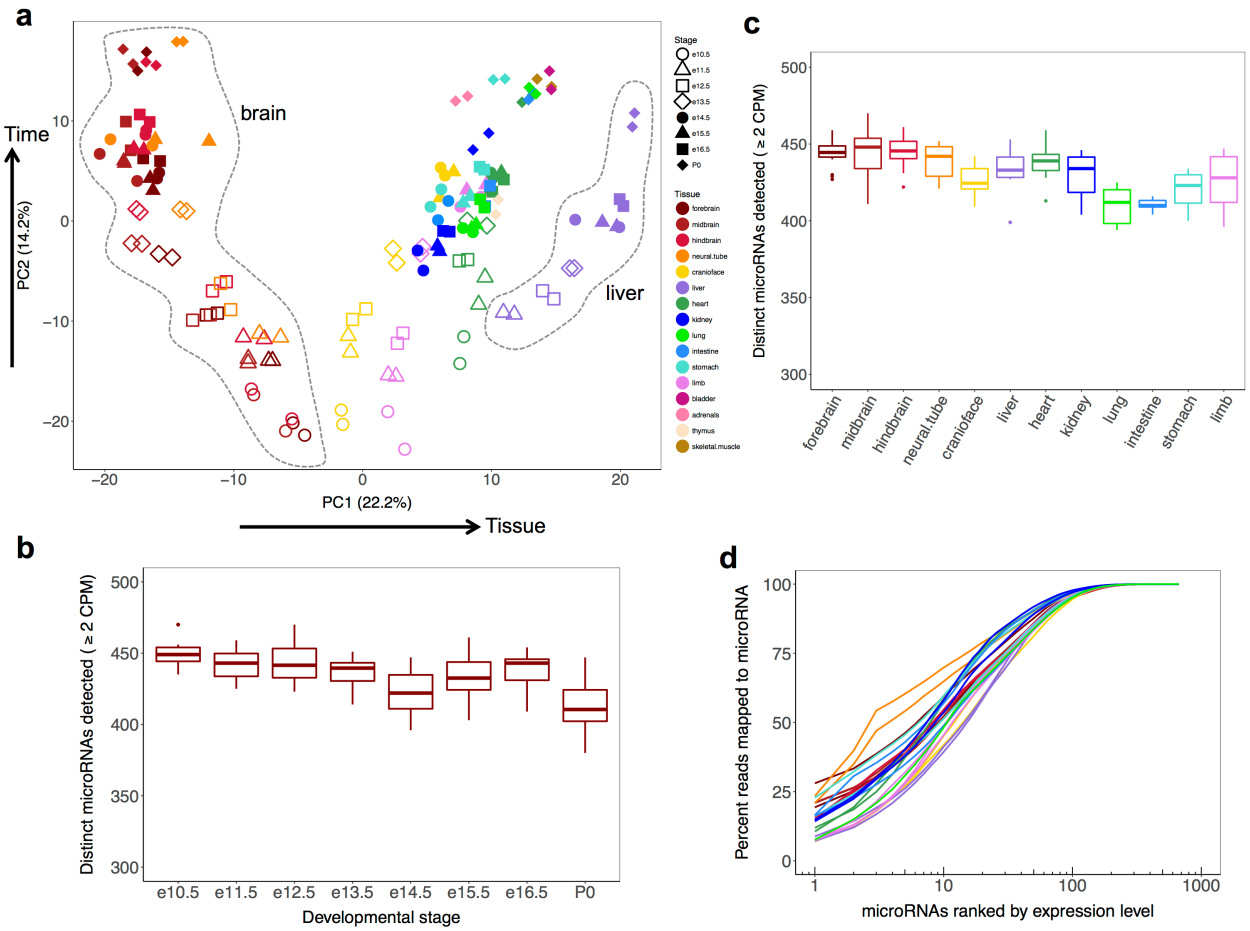
Figure 4.1: Overview of mouse and human ENCODE miRNA data sets.



(a) Primary miRNAs were profiled using mRNA-seq (> 200 nt) in human and mouse, pre-miRNAs and mature miRNAs were profiled in human using short RNA-seq (< 200 nt), and mature miRNAs in mouse were profiled using microRNA-seq (< 30 nt) and NanoString. (b) Primary tissues representative of major organ systems were profiled in mouse and human along various stages of embryonic and fetal development. (c) Comparison of normalized miRNA counts for GM12878, profiled using microRNA-seq and short RNA-seq, demonstrates high

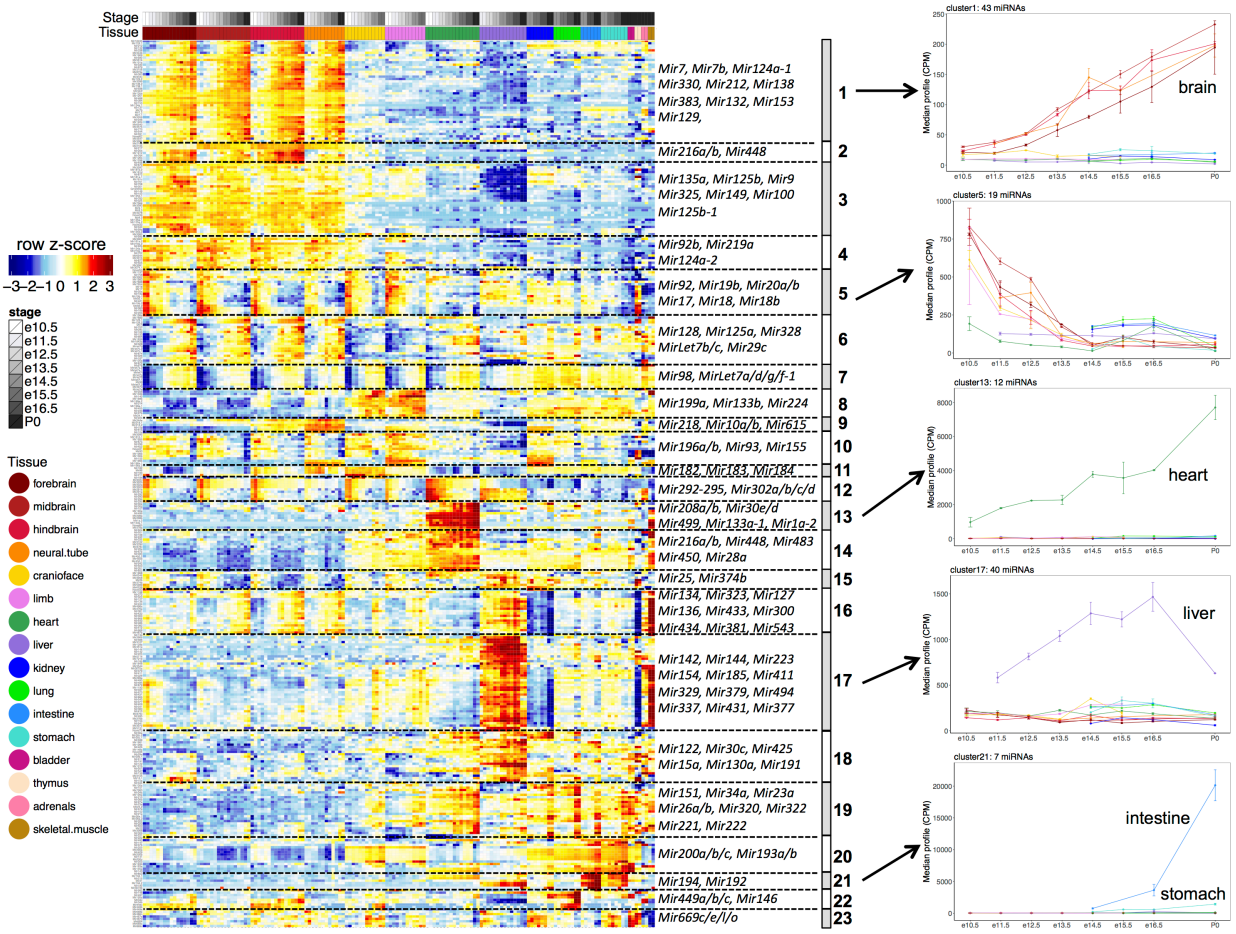
correlation between the two assays. (d) Heatmap of miRNA normalized counts in GM12878 and K562 cell lines shows that the samples cluster by cell type irrespective of profiling technique used or the date of sample preparation.

Figure 4.2: Global properties of mouse miRNA embryonic development time-course.



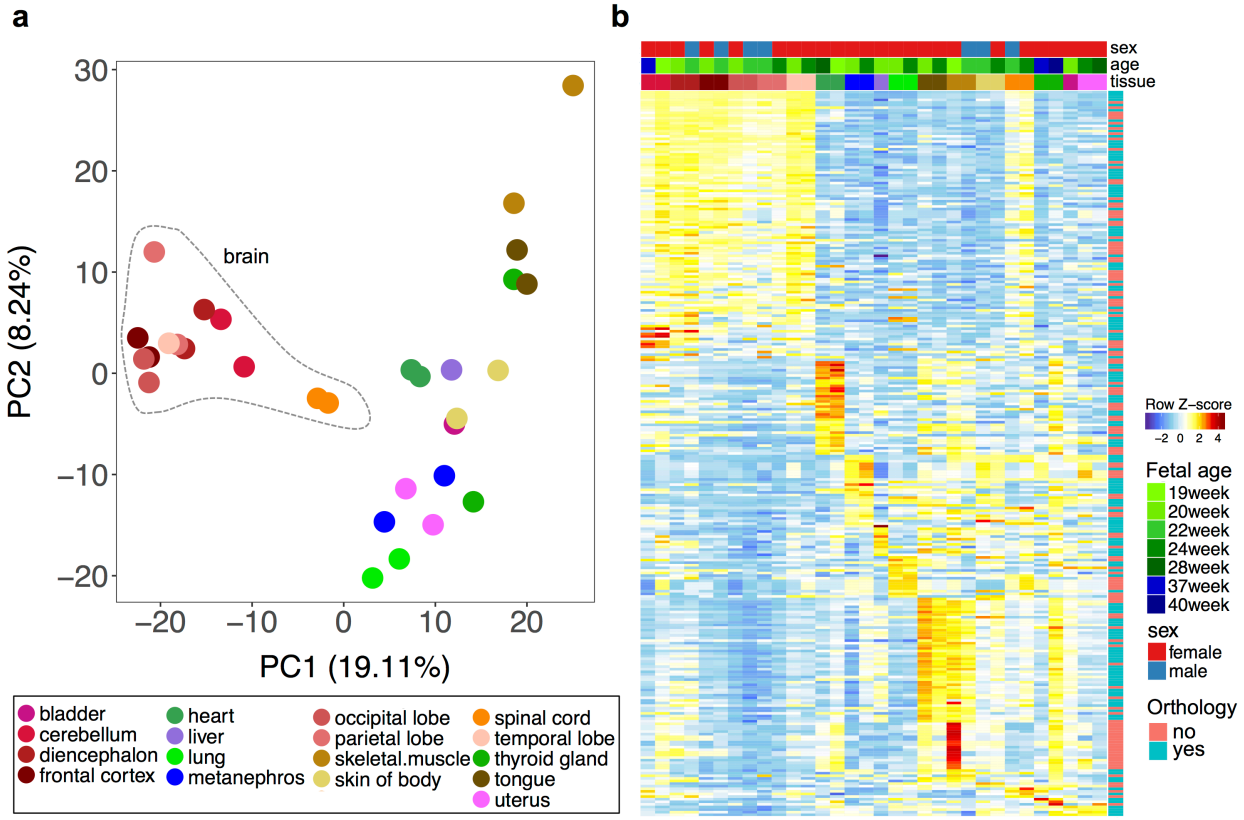
(a) Principal component analysis (PCA) of 16 mouse tissues across 8 developmental stages. Tissues are represented by specific colors while shapes denote the various developmental stages. (b) Number of distinct miRNAs detected in different developmental stages (minimum 2 CPM). (c) Number of distinct miRNAs detected in tissues (minimum 2 CPM). (d) Cumulative distribution of sequencing reads, accounted for by miRNAs ranked by expression levels, for tissues in developmental stage e14.5 that is representative of all stages. The top 100 miRNAs account for most of the miRNA sequencing reads.

Figure 4.3: Time series analysis of miRNAs across mouse embryonic development.



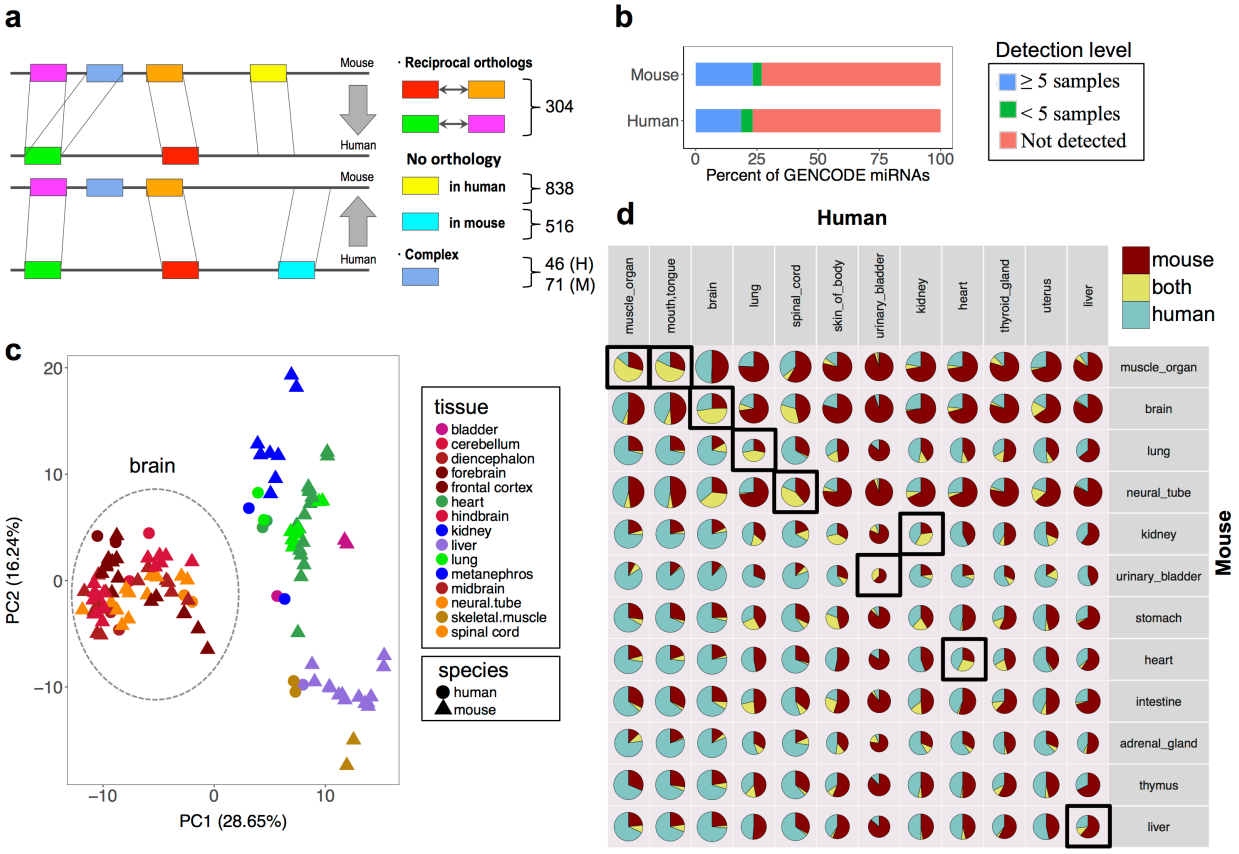
Heatmap representing the hierarchical clustering of miRNAs. The miRNAs shown were identified as differentially expressed by time-series analysis of 12 mouse tissues that were profiled in at least two embryonic development stages using the linear regression based algorithm maSigPro. The differentially expressed miRNAs were clustered into 23 non-redundant groups based on median expression profiles. 5 representative profiles of the miRNA clusters are included that exhibit tissue-specific and stage-specific dynamics.

Figure 4.4: Human fetal development miRNA transcriptome.



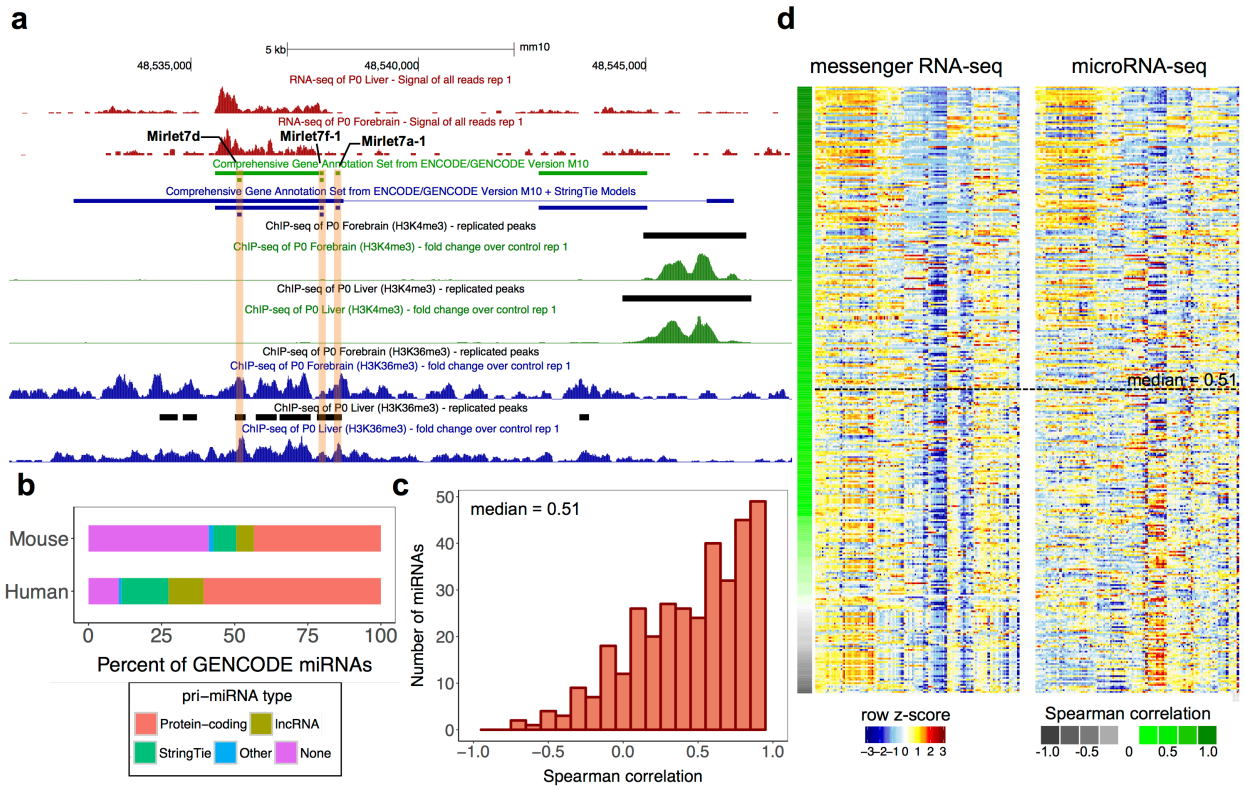
(a) Principal component analysis (PCA) of human tissues reveals distinct miRNA expression patterns in brain samples compared to other tissues. Colors denote different tissues. (b) Normalized expression levels of human tissue-specific miRNAs. Differential expression analysis reveals the largest set of differentially expressed miRNAs in brain and muscle samples.

Figure 4.5: Comparative dynamics of miRNAs during human and mouse development.



(a) 304 one-to-one orthologous human and mouse miRNAs were identified using reciprocal search. (b) The fraction of annotated miRNAs detected in mouse and human tissues (GENCODE v. M10 in mouse and GENCODE v. 25 in human). (c) Combined principal component analysis (PCA) of human and mouse samples. Triangles and circles denote mouse and human tissues respectively. Tissues are denoted by different colors. (d) Intersection of human and mouse tissue-specific miRNAs. For each pair of tissues in human and mouse we report the fraction of tissue-specific miRNAs in mouse only (red), human only (blue), or in both (yellow) within the 304 orthologous miRNAs. The sizes of the pie chart are in proportion to the numbers of tissue-specific miRNAs in the corresponding tissues.

Figure 4.6: Comparison of miRNAs and their primary transcripts using GENCODE annotations augmented with *ab initio* models.



(a) A genome browser snapshot of a representative *ab initio* gene model generated using mouse mRNA-seq data for 2 miRNAs (Mir-let7f-1 and Mir-let7a-1) that lack annotated GENCODE v. M10 pri-miRNAs. Additional evidence for the gene models is provided using H3K4me3 and K3K36me3 data in matching mouse samples. (b) Distribution of types of pri-miRNAs biotypes (protein-coding, lncRNA, *ab initio* gene models, and others) in mouse and human. Improvements in pri-miRNA annotations using the *ab initio* gene models are denoted by color green in both human and mouse. (c) Distribution of Spearman correlation (median = 0.51) among miRNAs and their corresponding pri-miRNAs. (d) Comparison of expression levels of pri-miRNAs and

their corresponding mature miRNAs. The rows are sorted with decreasing Spearman correlation top to bottom.

Figure 4.7: Distribution of Spearman correlations between microRNA-seq and NanoString for miRNAs included in the NanoString codeset.

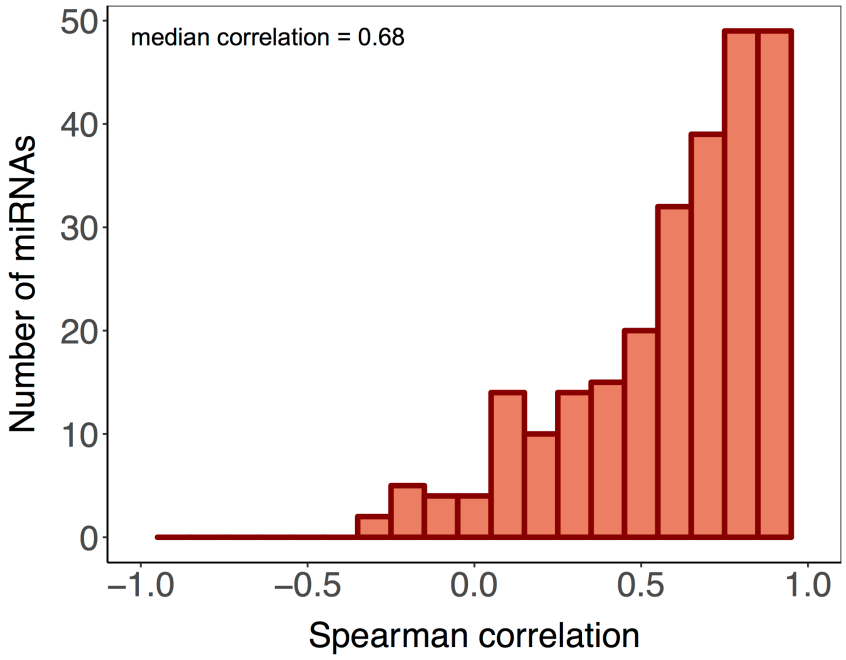
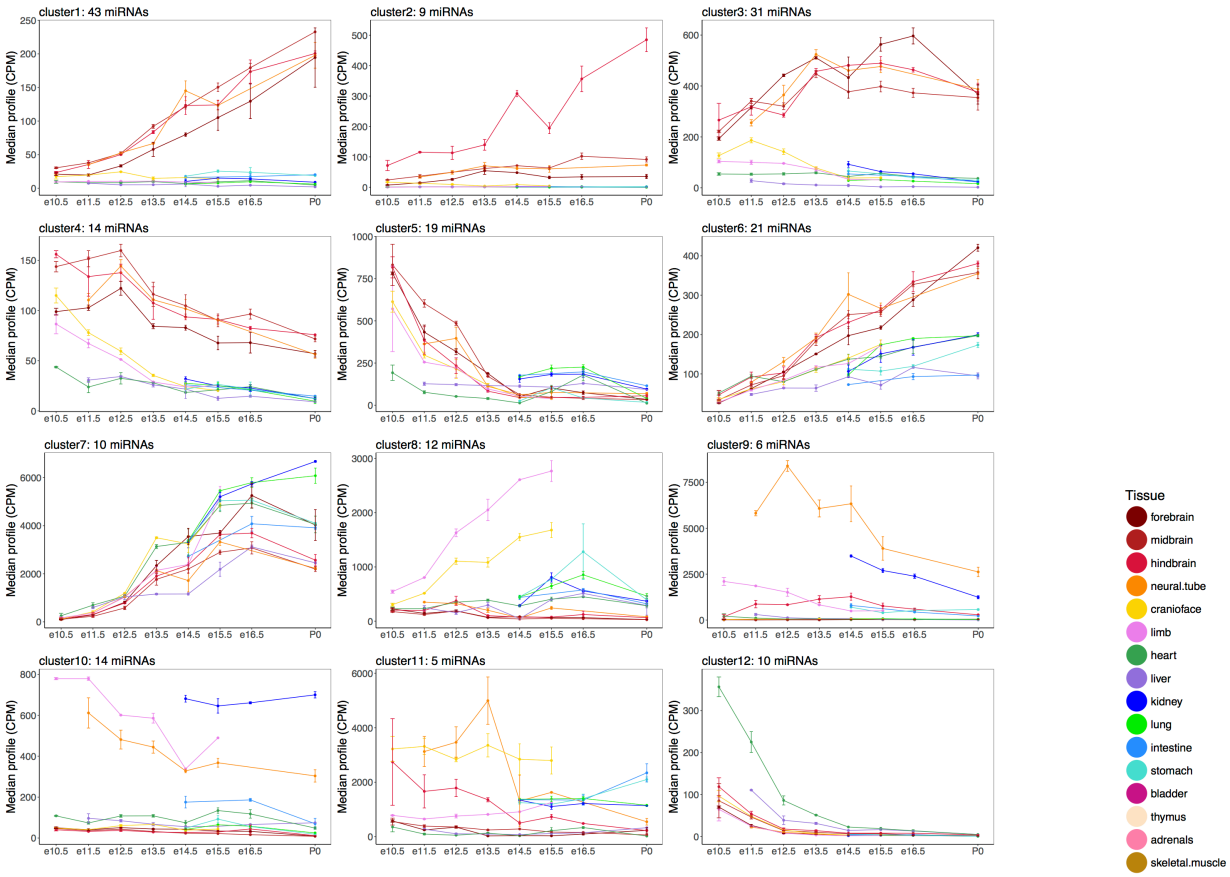
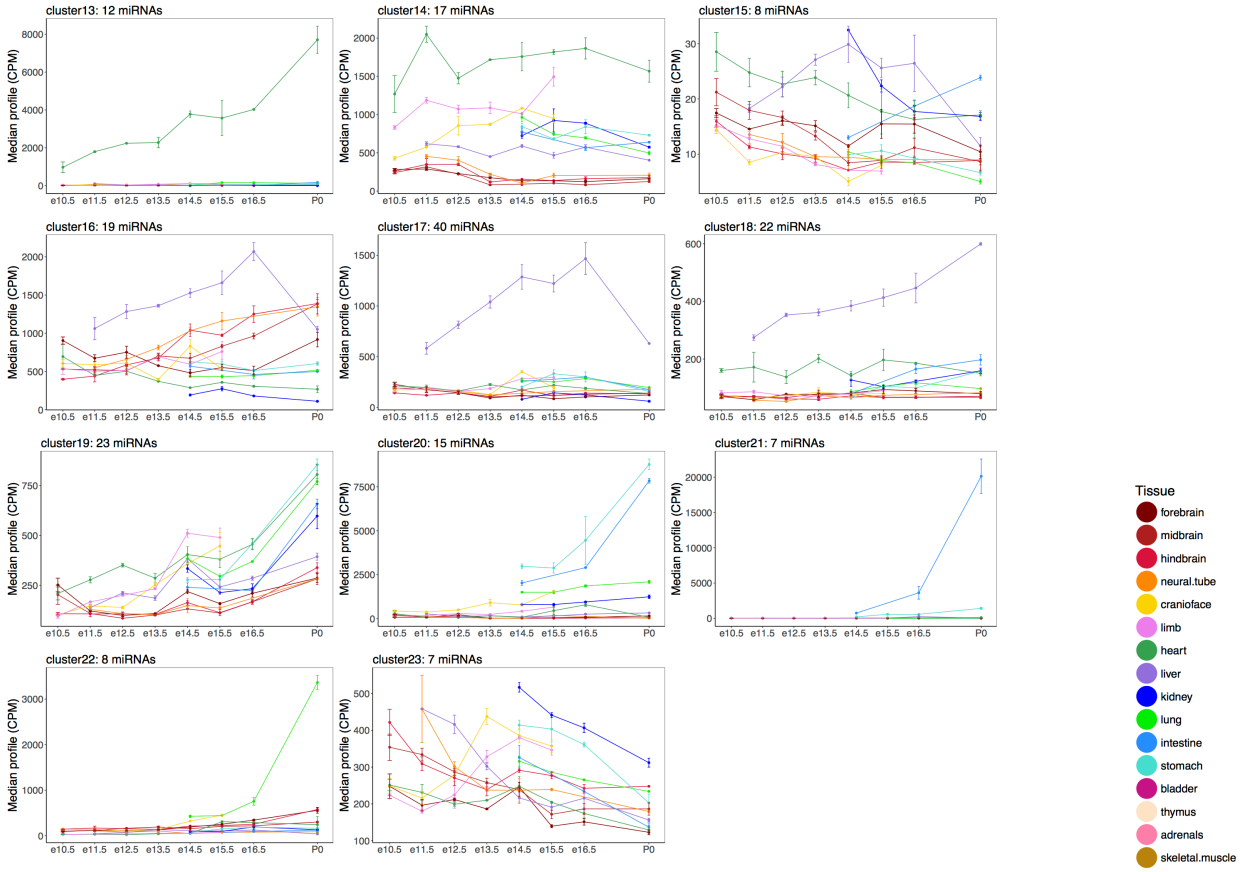


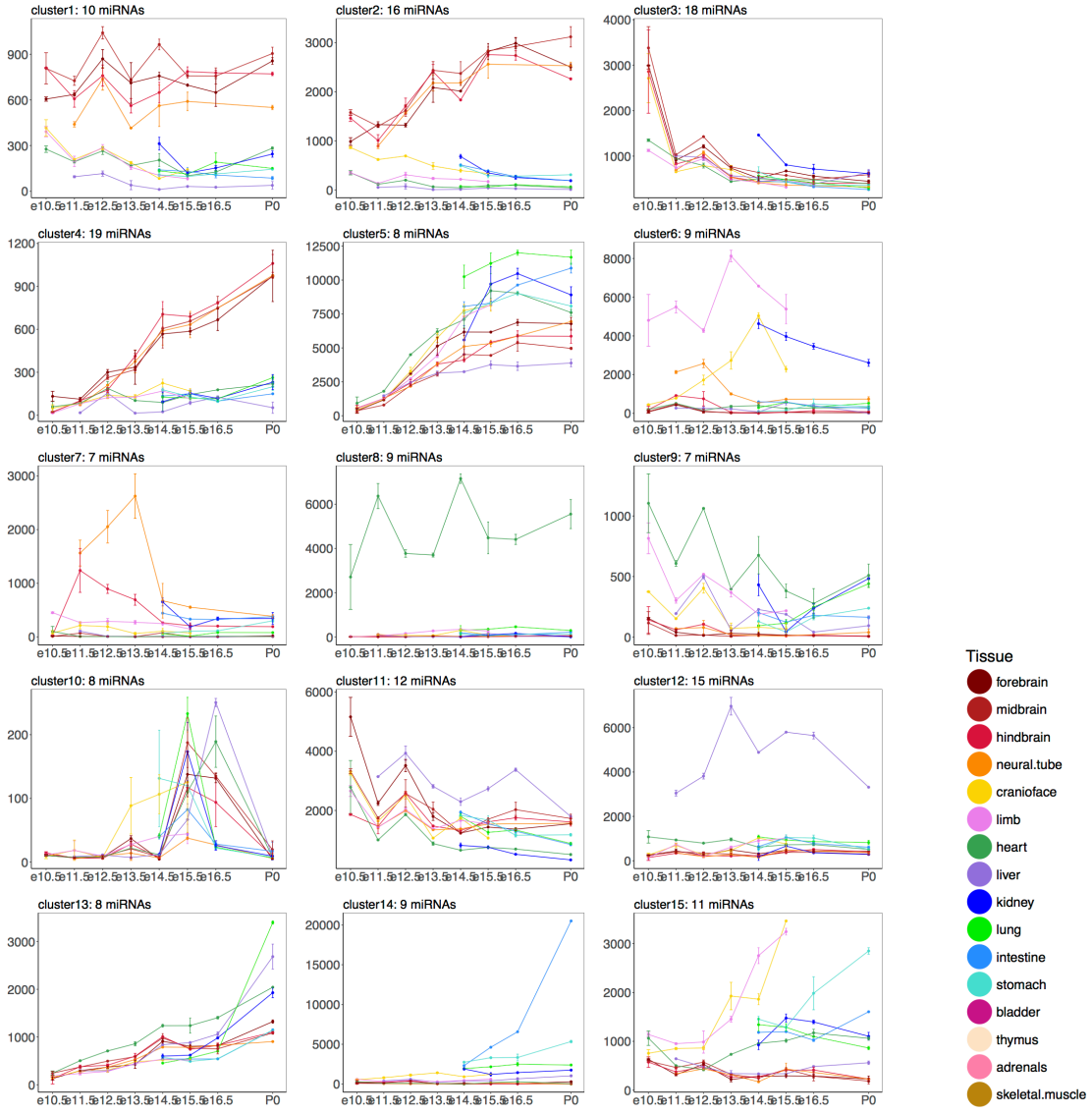
Figure 4.8: The median expression profiles of 23 clusters of miRNAs measured using microRNA-seq.





The median expression profiles of 23 clusters of miRNAs, measured using microRNA-seq, that were identified as differentially expressed by the linear regression based algorithm maSigPro.

Figure 4.9: The median expression profiles of 15 clusters of miRNAs, measured using NanoString.



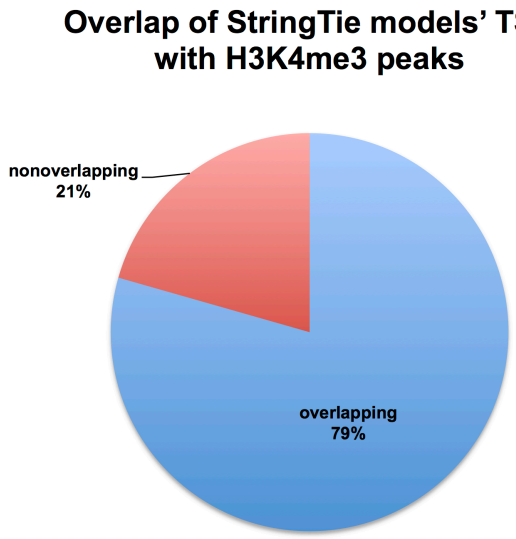
The median expression profiles of 15 clusters of miRNAs, measured using NanoString in matching samples, that were identified as differentially expressed by the linear regression based algorithm maSigPro**.**

Figure 4.10: Validation of miRNA expression with NanoString data.

NanoString/ miRNAseq	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	No match
1	2	1	3	2		1							1											0
2	4		6	1					1							1			1					2
3	1		1		8					2		3			1									2
4	5	1	1			6										2	1							3
5							5											1				2		0
6						1	1	5		2														0
7									3		4													0
8													5	4										0
9												2		4					1					0
10		1				1						1												5
11																4	3	1						4
12																2	10	3						0
13													1					1	4	1				1
14																				7	2			0
15								2						1				1	4	2				1

Overlap of miRNAs among the 15 clusters of differentially expressed miRNAs (left column) assayed using NanoString (Fig. 4.9) and the 23 clusters of differentially expressed miRNAs (top row) assayed with microRNA-seq (Fig. 4.8). Boxes with orange color indicate the clusters with the highest orthogonal overlap.

Figure 4.11: Proportion of overlap between the TSS of StringTie *ab initio* transcript models and the H3K4me3 peaks in matching samples.



The TSS of the transcript models were defined as ±250 bp regions from the 5'-end of the transcripts.

4.6 Methods and Materials

microRNA-seq from mouse embryonic tissues:

Mouse Embryonic Tissue Acquisition: A detailed protocol for tissue acquisition used for this study can be found at:

https://www.encodeproject.org/documents/631aa21c-8e48-467e-8cac-d40c875b3913/@@download/attachment/Tissue_Excision_Protocols_112414.pdf

RNA Isolation: Total RNA was obtained using mirVana miRNA isolation kit and protocol:

https://www.encodeproject.org/documents/f0cc5a7f-96a5-4970-9f46-317cc8e2d6a4/@@download/attachment/cms_055423.pdf

The protocol for genomic DNA removal:

https://www.encodeproject.org/documents/428a184d-7fa1-4599-9d8d-749c2eba7edd/@@download/attachment/cms_055740.pdf

Library Construction: The construction of microRNA-seq libraries was based on the previously published protocol (Roberts et al, 2015, Nucleic Acid Research) with some minor modifications listed below and without the highly abundant miRNA blocking step. Briefly, 500ng of total RNA with RIN (RNA integration number) higher than 9.0 was used as input material, together with spike-in control. 3'adapter was ligated to the sample with T4 RNA ligase 2, truncated (NEB), then reverse transcription primer was annealed to the 3'adapter in order to reduce the 5' and 3' adapter dimer. After that, 5' adapter was ligated to the product with T4 RNA ligase 1 (NEB). Here, we used a pool of four multiplex 5' adapters. At the end of the 5' adapter, there is a six-nucleotide spacer, which was present as the first six nucleotides in read 1 of the sequence data in

order to provide base diversity during the crucial first cycles. Ligation product was reverse transcribed with Superscript II (Invitrogen) and the cDNA was further amplified using Phusion high-fidelity PCR master mix (NEB). Primers used at the PCR stage introduce a barcode, used later for sample demultiplexing. PCR products were purified with Ampure XP beads (Beckman Coulter). To get rid of adapter-dimer and the other non-miRNA product, size selection of the microRNA-seq libraries was performed using 10% TBE-urea polyacrylamide gel (Bio-Rad) in hot (70C) TBE running buffer for 45 mins. The 140-nt denatured microRNA-seq library band was excised, eluted from the gel slice, precipitated by isopropanol and resuspended with 10ul EB buffer (QIAGEN). Library concentration was determined with Library Quantification Kit (KAPA Biosystems). The DNA Bioanalyzer assay is unable to show the accurate profile of the library and was not employed.

Sequencing: The microRNA-seq libraries were sequenced as 50 bp single-end reads on an Illumina HiSeq2000 sequencer.

short RNA-seq from human fetal tissues:

Detailed protocol for short RNA-seq library construction can be found at: https://www.encodeproject.org/ by searching each sample accession ID.

NanoString from mouse embryonic tissues:

The samples were prepared with NanoString human miRNA kit version 2.1 (based on miRBase v.18) following its protocol. In short, 100ng total RNA was used as starting material. Together with "spike-in" positive and negative controls, each target miRNA was ligated to a specific miRNAtag molecule and the chimeric miRNA:miRNAtag molecule was hybridized with

fluorescent-labeled probes overnight. The miRNA:miRNAtag chimeric molecule is long enough to ensure the efficiency and specificity of probe hybridization. After the samples were processed in NanoString nCounter PrepStation to remove unhybridized probes, they were immobilized and aligned in scanning cartridges and scanned in NanoString nCounter digital analyzer with the maximal resolution setting to achieve the counts of each individual target miRNA molecule recognized by probe.

Data processing and analysis:

Adapter trimming of mouse microRNA-seq and human short RNA-seq reads

Due to small size of miRNAs (<30 nt), adapter trimming of raw sequencing reads was an important step before mapping.

Mouse microRNA-seq read adapter trimming: We used Cutadapt v.1.7.1 with Python 2.7 to sequentially trim 5' and 3' adapters from raw reads. The 3' and 5' (a mixture of 4 sequences) adapter sequences are as follows:

3'_adapter_seq = "ACGGGCTAATATTTATCGGTGGAGCATCACGATCTCGTAT"

5'_adapter_seq1 = "^CAGTCG"

5'_adapter_seq2 = "^TGACTC"

5'_adapter_seq3 = "^GCTAGA"

5'_adapter_seq4 = "^ATCGAT"

cutadapt -a 3'_adapter_seq -e 0.25 --match-read-wildcards --untrimmed-output=$NO_3AD_FILE input.fastq | cutadapt -e 0.34 --match-read-wildcards --no-indels -m 15

-O 6 -n 1 -g 5'_adapter_seq1 –g 5'_adapter_seq2 -g 5'_adapter_seq3 -g 5'_adapter_seq4 --untrimmed-output=$NO_5AD_FILE --too-short-output = $TOO_SHORT_FILE - > trimmed_reads.fastq

Human short RNA-seq read adapter trimming: Reads were initially trimmed for TGGAATTCTC adapters and Ns with cutadapt with parameters: -m 16 --trim-n. In the case of polyA adapters, three additional parameters were used: -a A{10} -e 0.1 -n 10 (to iteratively remove longer polyA tails).

Mapping of mouse microRNA-seq reads:

Trimmed reads were mapped to the mouse genome (assembly mm10) with STAR v2.4.2a with parameters:
--runThreadN 16 --sjdbGTFfile ENCSR021VAV/ENCFF992LCK.gtf --alignEndsType EndToEnd --outFilterMismatchNmax 1 --outFilterMultimapScoreRange 0 --quantMode TranscriptomeSAM GeneCounts --outReadsUnmapped Fastx --outSAMtype BAM SortedByCoordinate --outFilterMultimapNmax 10 --outSAMunmapped Within --outFilterScoreMinOverLread 0 --outFilterMatchNminOverLread 0 --outFilterMatchNmin 16 --alignSJDBoverhangMin 1000 --alignIntronMax 1 --outWigType wiggle --outWigStrand Stranded --outWigNorm RPM

Mapping of human short RNA-seq reads:

Trimmed reads were mapped to the human genome (assembly hg38) with STAR v2.5.1b with parameters: --outFilterMultimapNmax 10 --outFilterMultimapScoreRange 0 --

outFilterScoreMinOverLread 0 --outFilterMatchNminOverLread 0 --outFilterMatchNmin 16 --outFilterMismatchNmax 1 --alignSJDBoverhangMin 1000 --alignIntronMax 1.

Finally, miRNA hairpins from GENCODE v25 were quantified by summing the reads that have 100% overlap with the hairpins.

Processing of mouse NanoString data

NanoString raw data was processed with nSolver Analysis Software version 2.5 from NanoString Technologies, Inc. Individual miRNA counts from each cell sample were normalized in order to correct for the variation of the applied RNA amount from different cell samples. The normalization factor was calculated using the geometric mean of top 100 expressed miRNA counts from the same sample. The normalized counts for all samples were further quantile normalized using Limma package in R/Bioconductor.

Differential analysis of human short RNA-seq data from human:

miRNA hairpins specific of a given tissue in human or in mouse were identified with the glmQLFit() and glmQLFTest() function() from the R package edgeR. We used a deviation coding system for contrasts which compares one tissue against all the others, with the function constr.sum(). P-values were adjusted for multiple testing with the Benjamini-Hochberg correction. Only hairpins with FDR<0.01 and a fold-change > 2 were considered tissue-specific.

Identification of orthologous miRNAs in mouse and human:

Orthologous miRNAs between human and mouse were identified through genomic alignment. Human miRNAs were lifted over to the mouse genome and vice versa, with a

minimum overlap requirement of 50%. If a human miRNA maps within 10kb of a mouse miRNA, and that mouse miRNA also maps within 10kb of the initial human miRNA, those miRNA are defined to be in reciprocal orthologous relationship.

Generation of ab initio transcripts models from mRNA-seq reads:

mRNA-seq reads were mapped to the mouse genome (assembly mm10) using STAR v.2.4.2a with the following parameters:

--genomeDir star.gencodeM10.index --readFilesIn $fastqs –sjdbGTFfile gencode.vM10.annotation.gtf --readFilesCommand zcat --runThreadN 8 --outFilterMultimapNmax 20 --alignSJoverhangMin 8 --alignSJDBoverhangMin 1 --outFilterMismatchNmax 999 --outFilterMismatchNoverReadLmax 0.04 --alignIntronMin 20 --alignIntronMax 1000000 --alignMatesGapMax 1000000 --outSAMunmapped Within --outFilterType BySJout --outSAMattributes NH HI AS NM MD XS --outSAMstrandField intronMotif --outSAMtype BAM SortedByCoordinate --sjdbScore 1

The alignments to the genome were assembled into ab initio transcripts using StringTie v.1.2.4 using the following parameters:

-G gencode.vM10.annotation.gtf -c 3 -p 8

The transcript models for each sample were merged into a single GTF file using StringTie with the following options:

stringtie --merge -G gencode.vM10.annotation.gtf -o merged.gtf -m 200 -F 1.0 -p 8

Single exon transcripts with no strand information were excluded. The expression levels of the GENCODE M10 and the new StringTie model transcripts were obtained using RSEM

v1.2.25 with the following parameters:

rsem-calculate-expression --star --star-path ~/STAR-STAR_2.4.2a/bin/Linux_x86_64/ -p 10 --gzipped-read-file fastqs RSEM_Index_GENCODE_M10_Plus_StringTieModels

Prediction of novel miRNAs from mouse microRNA-seq and human short RNA-seq reads:

Mouse: All the trimmed microRNA-seq reads of samples for the same tissue were pooled and novel miRNAs predicted using mirdeep2 v2.0.0.8. The parameters used were:

mapper.pl trimmed_reads.fastq -e –p mm10

miRDeep2.pl processed.reads.fastq mm10.fasta mapped.arf

We used miRBase v.21 mature and hairpin annotations for mouse and ENSEMBL v.85 rat hairpin annotations for the miRDeep2.pl step above. Novel miRNAs with a score of 4 or higher (corresponding to 70% or higher true positive confidence level), independently identified in at least two tissues, not overlapping GENCODE M10 annotated miRNAs and the genomic repeat regions, and expressed at 2 CPM minimum in at least one sample (both replicates) were kept for downstream analysis.

Time-series analysis of mouse microRNA-seq and NanoString data from mouse:

Time series analysis of the mouse microRNA-seq time-course was performed using maSigPro_1.42.0 in R 3.2.3. Briefly, each tissue (12 in total) that were assayed in two at least two developmental time points were analysed using degree 3 and maSigPro functions "p.vector(data, design = design.matrix, counts = TRUE)", "T.fit(p.vector_output, alfa = 0.01)", and "get.siggenes(T.fir_output, rsq=0.7, vars="all")". The number of clusters "k" obtained was

changed until 23 non-redundant clusters of genes were obtained using “see.genes(get$sig.genes, k = 23)”. The median profiles of the genes (standard errors denoted for the replicates) were plotted using ggplot2 package.

4.7 References

Bartel, D. P. (2004). MicroRNAs: Genomics, Biogenesis, Mechanism, and Function. *Cell*, *116*(2), 281–297. https://doi.org/10.1016/S0092-8674(04)00045-5

Cai, X., Hagedorn, C. H., & Cullen, B. R. (2004). Human microRNAs are processed from capped, polyadenylated transcripts that can also function as mRNAs. *RNA (New York, N.Y.)*, *10*(12), 1957–66. https://doi.org/10.1261/rna.7135204

Cao, D. D., Li, L., & Chan, W. Y. (2016). MicroRNAs: Key regulators in the central nervous system and their implication in neurological diseases. *International Journal of Molecular Sciences*, *17*(6), 1–28. https://doi.org/10.3390/ijms17060842

Ehrenreich, I. M., & Purugganan, M. (2008). MicroRNAs in plants: Possible contributions to phenotypic diversity. *Plant Signaling & Behavior*, *3*(10), 829–30. https://doi.org/10.1101/gad.1004402.of

Encode Consortium. (2012). An integrated encyclopedia of DNA elements in the human genome. *Nature*, *489*(7414), 57–74. https://doi.org/10.1038/nature11247

Fejes-Toth, K., Sotirova, V., Sachidanandam, R., Assaf, G. Hannon, G. J., Kapranov, P., Foissac, S., … Gingeras, T. R. (2009). Post-transcriptional processing generates a diversity of 5'-modified long and short RNAs. *Nature*, *457*(7232), 1028–1032. https://doi.org/10.1038/nature07759

Friedman, R. C., Farh, K. K. H., Burge, C. B., & Bartel, D. P. (2009). Most mammalian mRNAs are conserved targets of microRNAs. *Genome Research*, *19*(1), 92–105. https://doi.org/10.1101/gr.082701.108

Gao, Y., Schug, J., McKenna, L. B., Le Lay, J., Kaestner, K. H., & Greenbaum, L. E. (2011). Tissue-specific regulation of mouse MicroRNA genes in endoderm-derived tissues. *Nucleic Acids Research*, *39*(2), 454–463. https://doi.org/10.1093/nar/gkq782

Geiss, G. K., Bumgarner, R. E., Birditt, B., Dahl, T., Dowidar, N., Dunaway, D. L., … Dimitrov, K. (2008). Direct multiplexed measurement of gene expression with color-coded probe pairs. *Nature Biotechnology*, *26*(3), 317–25. https://doi.org/10.1038/nbt1385

Guo, Z., Maki, M., Ding, R., Yang, Y., Zhang, B., & Xiong, L. (2014). Genome-wide survey of tissue-specific microRNA and transcription factor regulatory networks in 12 tissues. *Scientific Reports*, *4*, 5150. https://doi.org/10.1038/srep05150

Han, J., Lee, Y., Yeom, K. H., Nam, J. W., Heo, I., Rhee, J. K., … Kim, V. N. (2006). Molecular Basis for the Recognition of Primary microRNAs by the Drosha-DGCR8 Complex. *Cell*, *125*(5), 887–901. https://doi.org/10.1016/j.cell.2006.03.043

He, L., & Hannon, G. J. (2004). MicroRNAs: small RNAs with a big role in gene regulation. *Nature Reviews. Genetics*, *5*(7), 522–531. https://doi.org/10.1038/nrg1415

Herrero, J., Muffato, M., Beal, K., Fitzgerald, S., Gordon, L., Pignatelli, M., … Flicek, P. (2016). Ensembl comparative genomics resources. *Database*, *2016*, 1–17. https://doi.org/10.1093/database/bav096

Kozomara, A., & Griffiths-Jones, S. (2011). MiRBase: Integrating microRNA annotation and deep-sequencing data. *Nucleic Acids Research*, *39*(SUPPL. 1), 152–157. https://doi.org/10.1093/nar/gkq1027

Ladewig, E., Okamura, K., Flynt, A. S., Westholm, J. O., & Lai, E. C. (2012). Discovery of hundreds of mirtrons in mouse and human small RNA data. *Genome Research*, *22*(9), 1634–1645. https://doi.org/10.1101/gr.133553.111

Lagos-Quintana, M., Rauhut, R., Yalcin, A., Meyer, J., Lendeckel, W., & Tuschl, T. (2002).

Identification of tissue-specific MicroRNAs from mouse. *Current Biology*, *12*(9), 735–739. https://doi.org/10.1016/S0960-9822(02)00809-6

Lee, R. C., Feinbaum, R. L., & Ambros, V. (1993). The C. elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14. *Cell*, *75*(5), 843–854. https://doi.org/10.1016/0092-8674(93)90529-Y

Lin, S., & Gregory, R. I. (2015). MicroRNA biogenesis pathways in cancer. *Nature Review Cancer*, *15*(6), 321–333. https://doi.org/10.1038/nrc3932

Ludwig, N., Leidinger, P., Becker, K., Backes, C., Fehlmann, T., Pallasch, C., … Keller, A. (2016). Distribution of miRNA expression across human tissues. *Nucleic Acids Research*, *44*(8), 3865–3877. https://doi.org/10.1093/nar/gkw116

Mestdagh, P., Hartmann, N., Baeriswyl, L., Andreasen, D., Bernard, N., Chen, C., … Vandesompele, J. (2014). Evaluation of quantitative miRNA expression platforms in the microRNA quality control (miRQC) study. *Nature Methods*, *11*(8), 809–815. https://doi.org/10.1038/nmeth.3014

Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L., & Wold, B. (2008). Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nature Methods*, *5*(7), 621–628. https://doi.org/10.1038/nmeth.1226

Ninova, M., Ronshaugen, M., Griffiths-jones, S., & Griffiths-jones, S. a M. (2014). Fast-evolving microRNAs are highly expressed in the early embryo of Drosophila virilis. *Rna*, 360–372. https://doi.org/10.1261/rna.041657.113.

Roberts, B. S., Hardigan, A. A., Kirby, M. K., Fitz-Gerald, M. B., Wilcox, C. M., Kimberly, R. P., & Myers, R. M. (2015). Blocking of targeted microRNAs from next-generation sequencing libraries. *Nucleic Acids Research*, *43*(21), 1–8. https://doi.org/10.1093/nar/gkv724

Romaine, S. P. R., Tomaszewski, M., Condorelli, G., & Samani, N. J. (2015). MicroRNAs in cardiovascular disease: an introduction for clinicians. *Heart (British Cardiac Society)*, *101*(12), 921–8. https://doi.org/10.1136/heartjnl-2013-305402

Vidigal, J. A., & Ventura, A. (2015). The biological functions of miRNAs: Lessons from in vivo studies. *Trends in Cell Biology*, *25*(3), 137–147. https://doi.org/10.1016/j.tcb.2014.11.004

Wienholds, E., Kloosterman, W. P., Miska, E., Alvarez-saavedra, E., Berezikov, E., Bruijn, E. De, … Plasterk, R. H. (2005). MicroRNA Expression in Zebrafish Embryonic Development. *Science*, *309*(July), 310–311.

Wyman, S. K., Knouf, E. C., Parkin, R. K., Fritz, B. R., Lin, D. W., Dennis, L. M., … Tewari, M. (2011). Post-transcriptional generation of miRNA variants by multiple nucleotidyl transferases contributes to miRNA transcriptome complexity. *Genome Research*, *21*(9), 1450–1461. https://doi.org/10.1101/gr.118059.110

Yue, F., Cheng, Y., Breschi, A., Vierstra, J., Wu, W., Ryba, T., … Ren, B. (2014). A comparative encyclopedia of DNA elements in the mouse genome. *Nature*, *515*(7527), 355–64. https://doi.org/10.1038/nature13992

Zeng, W., Jiang, S., Kong, X., El-ali, N., Ball, A. R., Ma, C. I.-H., … Mortazavi, A. (2016). Single-nucleus RNA-seq of differentiating human myoblasts reveals the extent of fate heterogeneity. *Nucleic Acids Research*, 1–13. https://doi.org/10.1093/nar/gkw739

Zhao, W., Zhao, S.-P., Zhao, Y.-H., Zhao, W., Zhao, S.-P., & Zhao, Y.-H. (2015). MicroRNA-143/-145 in Cardiovascular Diseases. *BioMed Research International*, *2015*, 1–9. https://doi.org/10.1155/2015/531740

Chapter 5

Future Directions

Head regeneration in Hydra is one of the most widely studied developmental processes in cnidarians. It involves the reestablishment of the head organizer at the apical end when the hypostome (containing the head organizer) is removed by bisection. The formation of a bud, the normal asexual mode of reproduction of Hydra, is also initiated by formation of a head organizer at the mid-body location where a bud arises. Wnt signaling is known to be involved in both processes (Hobmayer et al., 2000). We performed time-courses of gene expression in both processes to reveal their genome-wide transcriptome similarities and differences. Whether the head organizer plays a role in embryonic development in Hydra is not known. Although it is most likely also involved in the development of the structure of the animal during embryogenesis. Future extensions of this study to the comparison with the head organizer formation during embryogenesis will reveal the extent of reuse of the normal developmental program during head regeneration. Furthermore, the RNA-seq experiment in this study has shed light on genome-wide gene expression patterns during formation of the head organizer in Hydra and was done on large sections from regenerating heads or buds to collect enough RNA for normal RNA-seq libraries. With the advent of single-cell RNA-seq techniques, measuring the gene expression of the apical most cells during regeneration of individual hydra will provide a better resolution of changes in gene expression. The head organizer in hydra is estimated to consist of 50-300 cells at the apical tip of the head. Single-cell profiling at a greater temporal resolution should provide additional insights into the role of signaling pathway and the processes that initiate and maintain it.

Comparative studies of metazoans separated by long evolutionary distances are interesting from an evolutionary developmental biology perspective. Such studies have provided important insights into the conservation of developmental programs as well as loss or gain of

new programs leading to the evolution of morphologies and functions. Along the long route of metazoan evolution, occasional sudden changes have led to drastic divergence among metazoans. An example of this drastic change occurred about 600 million years when cnidarians (with simpler body plans consisting of radial symmetry) and bilaterians (with left-right symmetry) diverged. What genetic changes led to this divergence? From recent genome sequencing efforts of cnidarians (Chapman et al., 2010) (Putnam et al., 2007), we now know that the divergence of cnidarians and bilaterians was most likely not due to changes in gene content. Evolutionary innovations in gene regulatory mechanisms have been postulated to effect metazoan evolution. Towards answering such as question we mapped the *cis*-regulatory landscape in Hydra during the well-studied process of head regeneration. In our work described in Chapter 3, we determined a list of 3018 candidate enhancer-like elements in the Hydra genome that contain chromatin accessibility and histone modification characteristics consistent with those of the bilaterian enhancers. Previously over 5000 enhancers were reported in another cnidarian *Nematostella vectensis* using histone modification evidence as well (Schwaiger, 2014). In case of Hydra, our candidate enhancer-like elements are at least 2 Kilobases (Kb) away from the nearest annotated transcription start site. It seems that cnidarians are similar to bilaterians in terms of the genomic architecture of enhancers except one important difference. In most of the bilaterians, once the right combination of transcription factors are bound to the transcription factor biding sites (TFBSs) at an enhancer, the enhancer region is brought within close proximity of the target gene promoter by DNA looping mediated by the transcription factor (TF) CTCF (Bulger & Groudine, 1999). The CTCF gene is missing in *Nematostella vectensis* (Heger, Marin, Bartkuhn, Schierenberg, & Wiehe, 2012) and most likely not present in Hydra genome as well since we found no evidence for it based on Blast search of its annotated genes. A likely

mechanism of enhancer function in cnidarian genomes could be through non-CTCF mediated DNA looping. An important future avenue for research into gene regulation in cnidarians is determining the mode of enhancer function. Furthermore, reporter assays for validating the set of enhancers and their spatio-temporal activities in Hydra will help in better understating their roles in Hydra biology.

We have generated a near-comprehensive atlas of microRNA (miRNA) expression in tissues and organs representative of all major organ systems in mouse during a time-course of embryonic development. miRNAs fine-tune the expression of their target genes (Bartel, 2009) and an important area of understanding their roles in development is determining their targets. Simultaneous profiling of miRNAs' and the expression levels of messenger RNA in the same cells and tissues can be useful in inferring miRNA-mRNA target relationships from partial correlations in their expression patterns and understand how networks of microRNA regulate developmental processes during embryonic development.

References

Bartel, D. P. (2009). MicroRNAs: Target Recognition and Regulatory Functions. *Cell*, *136*(2), 215–233. https://doi.org/10.1016/j.cell.2009.01.002

Bulger, M., & Groudine, M. (1999). Looping versus linking: Toward a model for long-distance gene activation. *Genes and Development*, *13*(19), 2465–2477. https://doi.org/10.1101/gad.13.19.2465

Chapman, J. A., Kirkness, E. F., Simakov, O., Hampson, S. E., Mitros, T., Weinmaier, T., … Steele, R. E. (2010). The dynamic genome of Hydra. *Nature*, *464*(7288), 592–596. https://doi.org/10.1038/nature08830

Heger, P., Marin, B., Bartkuhn, M., Schierenberg, E., & Wiehe, T. (2012). The chromatin insulator CTCF and the emergence of metazoan diversity. *Proceedings of the National Academy of Sciences*, *109*(43), 17507–17512. https://doi.org/10.1073/pnas.1111941109

Hobmayer, B., Rentzsch, F., Kuhn, K., Happel, C. M., von Laue, C. C., Snyder, P., … Holstein, T. W. (2000). WNT signalling molecules act in axis formation in the diploblastic metazoan Hydra. *Nature*, *407*(6801), 186–189. https://doi.org/10.1038/35025063

Putnam, N. H., Srivastava, M., Hellsten, U., Dirks, B., Chapman, J., Salamov, A., … Rokhsar, D. S. (2007). Sea Anemone Genome Reveals Ancestral Eumetazoan Gene Repertoire and Genomic Organization. *Science*, *317*(5834), 86–94. https://doi.org/10.1126/science.1139158

Schwaiger, M. (2014). Evolutionary conservation of the eumetazoam gene regulatory landscape - Supplemental Figures. *Genome Research*, 1–13. https://doi.org/10.1101/gr.162529.113.Freely