THE IMPACT OF THE MICROBIOTA ON TRANSCRIPTIONAL REGULATION IN THE VERTEBRATE INTESTINE

J. Grayson Camp

A dissertation submitted to the faculty of the University of North Carolina at Chapel Hill in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Curriculum of Genetics and Molecular Biology

Chapel Hill
2012

Approved By:

John F. Rawls, Ph.D.

Scott J. Bultman, Ph.D.

Jason D. Lieb, Ph.D.

P. Kay Lund, Ph.D.

William F. Marzluff, Ph.D.

Praveen Sethupathy, Ph.D.

ABSTRACT

J. GRAYSON CAMP: The Impact of the microbiota on transcriptional regulation in the vertebrate intestine
(Under the direction of John F. Rawls)

Animals evolved in a world pre-dominated by microscopic organisms. Colonization of intestinal tracts at birth by microbes initiates the next generation of an ancient symbiosis that profoundly impacts our physiology and pathophysiology. A record of this symbiosis is encoded in our genomes. In this dissertation, I explore how regulatory regions embedded in non-genic DNA mediate transcriptional responses to the intestinal microbiota. Extensive research has demonstrated that the complex community of microbes residing within our intestine (the gut microbiota) contributes biochemistries that enhance nutrient digestion, metabolize xenobiotics, and collectively function as an important environmental factor that modulates host energy balance and immunity. However, the mechanisms that host cells use to perceive and respond to these microbial activities are not well understood. I used the zebrafish and mouse gnotobiotic models to define mechanisms by which the microbiota regulates host transcription in the intestinal epithelium at the single gene and genome-wide scales. The intestinal microbiota enhances dietary energy harvest leading to increased lipid storage in peripheral tissues. This effect is caused in part by the microbial suppression of intestinal expression of a circulating inhibitor of Lipoprotein lipase called Angiopoietin-like 4 (Angptl4/Fiaf). I utilized the zebrafish in which host regulatory DNA can be rapidly analyzed in a live, transparent, and gnotobiotic vertebrate to define the *cis*-regulatory mechanisms controlling *angptl4* transcription. I discovered an intronic *cis*-regulatory module (CRM)

that confers intestine-specific transcription and microbial suppression of *angptl4*. I used comparative sequence analysis from 12 fish species, functional mapping, and mutagenesis to define the minimal set of regulatory sequences required for activity of the *angptl4* intestinal CRM. I applied computational prediction and DNA affinity chromatography to discern candidate transcription factors regulating *angptl4* intestinal expression. At the genomic level, I employed DNase-seq and FAIRE-seq in the intestine of germ-free and conventionally-raised mice and zebrafish to facilitate the discovery of CRMs meditating host responses to the microbiota genome-wide. This work provides a novel paradigm for understanding how microbial signals interact with tissue-specific regulatory networks to control host gene expression and elucidates mechanisms mediating over 500 million years of co-existence and co-evolution of vertebrate hosts with their intestinal microbiota.

ACKNOWLEDGEMENTS

Foremost, I would like to thank my mentor, John Rawls, for all of his guidance and patience during my graduate training. His curiosity, keen sense for the proper experiment, and remarkable ability to see the bigger picture has been a rich source of inspiration for me and I hope to continue to learn from him as I mature as a scientist. His steadfast support and appreciation for both the scientific and personal well being of the individuals in his lab has made this journey a great experience. I also thank him for turning me on to Parasite Rex and my interview with him for IBMS, as this marked a node in my life.

I would like to thank everyone in the Rawls Lab, past and present: Michelle Kanther, Ivana Semova, Ed Flynn, James Minchin, Jordan Cocchiaro, Sandi Wong, Jim Davison, Amy Jazwa, Chad Trent, Lantz Mackey, Laura Mackey, Linh Pham, Jessica Russell. These people have contributed to my life in different ways and I appreciate the friendships and common experiences we've had together over the past 5 years. Michelle Kanther was a model for her organizational skills and did her best to try to keep me in line. Ivana Semova was a thought provoking lab mate and friend and we had many great times together in and out of lab. Jordan Cocchiaro was a conscientious lab mom and fellow craft enthusiast. James Minchin provided me with practical knowledge about figure making, zebrafish rearing, cloning and from time to time a nice hug. Special thanks also to Ed Flynn who kept the lab running smoothly and always had a positive

attitude. Amy Jazwa was my first mentee and I learned many valuable lessons from this experience.

The research environment at UNC has been an open, inspiring, and collaborative place to do science. My committee members provided direction and much appreciated advice on my projects. I was lucky enough to collaborate with the labs of Jason Lieb and Greg Crawford and appreciate especially the invaluable technical assistance of Jeremy Simon, Chris Frank, and Yoichiro Shibata. Special thanks to Sausyty Hermreck and Cara Marlow for their immense support, and to Bob Duronio and Pat Brennwald for their leadership. I was positively impacted by many more fellow scientists and research staff, too many to mention by name, but I am grateful to all of them.

The support of my friends and family has been very important to me during my time at UNC, and it is difficult to express how much I am thankful for them all. Maggie McCormick is like a sister and will understand why I don't write more about her contribution to my well being. Agos Santoro is a great friend and being around her lively personality has made me happy over the past couple of years. Doug McIntyre was an ideal roommate. I heartily enjoyed listening to his ponderings and appreciate his sense of Walker Texas Ranger-like ideals. Anne-Marie Neiser exemplified honesty, moralty, and optimism and helped me grow up. Special thoughts go out to Robert Sons, Bryan Richardson, Chris Schmidt, Meghan Morgan, all of the fellow IBMS students, and my Roberson street roommates that made my early graduate career a great experience. Thanks to Garret Smith for loving the Avett Brothers and Josh Ritter. Colin Lickwar is a deep and creative thinker and doer and our conversations expanded my range of thought. My Mom, Dad, and Sister have been unflinching sources of love. They helped make me who I am and I hope that they can forgive my neglect over the past five years. Thanks also to my brother in law Mike for his curiosity and his help in making my niece. Ich möchte meine Tanzpartnerin Bäbel danken for her unique combination of love, will,

intelligence, and exuberance. It would be difficult to imagine my life without her to help me travel through it.

At the end of each research chapter I have acknowledged the specific contributions that other people have made to the work presented in that chapter. The results presented in Chapter 3 of this dissertation derive from a previously published article:

J. Gray Camp, Amy L. Jazwa, Chad M. Trent, and John F. Rawls. Intronic *cis*-regulatory modules mediate tissue-specific and microbial suppression of *Angptl4/Fiaf* transcription. *PLoS Genetics* 8: e1002585.

In Chapter 5, Chris Frank generated the ileum libraries and helped with the computational analysis of the data. Yoichiro Shibata aligned the DNase-seq reads to the mouse genome and helped with peak calling. Jeremy Simon performed the initial processing of the zebrafish FAIRE-seq reads, aligned them to the zebrafish genome, and performed the peak calling. The liver and kidney DNase datasets are unpublished and were generously provided by the Crawford lab. John Mayfield dissected the liver and kidney tissue from the mice and Alexias Safi made the DNase libraries.

Unless otherwise noted, I designed, implemented the methodologies, interpreted the results, and wrote the manuscript, all in concert with my advisor John Rawls.

PREFACE

The Waking
By Theodore Roethke

I wake to sleep, and take my waking slow.
I feel my fate in what I cannot fear.
I learn by going where I have to go.

We think by feeling. What is there to know?
I hear my being dance from ear to ear.
I wake to sleep, and take my waking slow.

Of those so close beside me, which are you?
God bless the Ground! I shall walk softly there,
And learn by going where I have to go.

Light takes the Tree; but who can tell us how?
The lowly worm climbs up a winding stair;
I wake to sleep, and take my waking slow.

Great Nature has another thing to do
To you and me, so take the lively air,
And, lovely, learn by going where to go.

This shaking keeps me steady. I should know.
What falls away is always. And is near.
I wake to sleep, and take my waking slow.
I learn by going where I have to go.

TABLE OF CONTENTS

List of Figures

List of Tables

CHAPTER 1

Introduction

The body surfaces of humans and other animals are colonized at birth by microorganisms. The majority of microbial residents on the human body exist within gastrointestinal tract (GI) communities, where they engage in symbiosis with host cells. The host genome encodes the ability to respond to microbial activities and therefore constitutes a nexus and historical record for this ancient symbiosis. Gene-specific and genome-wide profiling of host gene expression has provided an important window into the microbial impact on host physiology and pathobiology, however the mechanisms underlying host transcriptional responses to the microbiota are poorly understood. Recent advances in high-throughput sequencing have expanded our ability to perceive the membership and physiologic traits of microbial communities along the GI tract. These same tools have in parallel dramatically expanded the functional understanding of vertebrate genomes in the fields of comparative genomics, transcriptional regulation, and chromatin biology. I propose in this dissertation that it is time to merge microbiota research with these fields in order to gain new mechanistic insights into how microbial symbiosis can impact transcriptional regulation and its evolution on a genomic scale. The following chapters will describe our current understanding of the microbial impact on transcriptional regulation in the intestinal epithelium, the contributions this dissertation provides to advancing that knowledge, and strategies to further probe the ancient relationship between our own cells and those of our microbial counterparts. Chapter 2 serves as a primer on host-microbiota symbioses and provides motivation for developing

methods to study transcriptional regulation and genomics in the vertebrate intestinal epithelium. Chapter 3 is a gene-centric analysis of the mechanisms that control transcription of *angiopoietin-like 4 (angptl4*) a gene that is expressed in the intestinal epithelium, functions in systemic lipid metabolism, and is dynamically regulated by the microbiota. I utilized the unique features of the zebrafish model to elucidate and characterize the *in vivo* activity of multiple DNA *cis*-regulatory modules (CRMs) that confer tissue-specific expression and microbial control of *angptl4*. The functional counterparts to CRMs are protein factors that bind DNA in *trans* to specify a genomic locus for transcriptional activation or repression. In Chapter 4, I discuss my efforts towards discovering transcription factors that regulate intestinal expression of zebrafish *angptl4*. This focused analysis on a single gene fostered an in depth appreciation of the intricacies that underlie gene expression programs and also established a set of methods to functionally assay CRMs *in vivo* in the presence and absence of a microbiota. Further, this work highlighted that sequence alignment alone limits the discovery of active regulatory regions in the zebrafish and other genomes. My ultimate goal is to understand how host-microbiota symbiosis impacts the evolution of non-coding *cis*-regulatory DNA. I therefore expanded my efforts to the genomic scale. In Chapter 5, I elucidated the genome-wide regulatory map of open chromatin in mouse and zebrafish intestinal epithelial cells and set forward plans to probe the impact of the microbiota on this chromatin landscape. Finally, in Chapter 6, I discuss future research initiatives at the interface of host-microbiota symbioses and transcriptional genomics. Cumulatively, this body of work vertically advanced the field of lipid metabolism by providing novel molecular mechanisms for tissue-specific and microbial regulation of *angptl4* expression, and provides a powerful new platform for genome-wide discovery and characterization of *cis/trans* regulatory programs mediating host-microbiota symbioses.

CHAPTER 2

Host-Microbiota Symbiosis and Transcription in the Vertebrate Intestine

2.1 Overview

Vertebrate gastro-intestinal (GI) tracts are home to a vast community of microorganisms that are integral for the development and health of the host animal. This chapter introduces salient features of host-microbe symbiosis highlighting the broad impact of the microbiota on intestinal epithelial cell biology in the small intestine. I provide an overview of current knowledge of transcriptional regulation in the intestinal epithelium and discuss how existing genomic approaches can be applied to elucidate regulatory DNA and transcription factors mediating intestine-specific responses to the microbiota. I discuss ways in which integrating genomic views of transcription with microbiota research will yield novel insight into the impact of environmental factors on transcription regulatory programs and genome evolution. Finally, I reinforce the importance of further developing gnotobiotic model systems to explore and functionally test predictions generated by genome-wide datasets *in vivo*.

2.2 Introduction

Multicellular animals (metazoans) evolved in a world that was predominated for billions of years by single-celled organisms. The advent of multicellularity [1] in the pre-metazoan lineages approximately 700-800 million years ago [2] led to a remarkable set of evolutionary innovations allowing for increased cellular specialization and organismal growth. This growth created new physical habitats and metabolic niches for intrepid microorganisms to colonize, and gave rise to anatomically distinct symbiotic relationships between present day animals and associated microbes. The gastrointestinal (GI) tract of vertebrates and other animals is one of the most diverse habitats colonized by microbes (known as the intestinal or gut microbiota or flora). Microbes reach high densities within the GI tract of animals where they impact various aspects of host biology including nutrient digestion [3], xenobiotic metabolism [4], epithelial barrier function [5], immune homeostasis [6,7] and collectively function as an important environmental factor that modulates host energy storage [8-11]. It is no surprise then that the intestinal microbiota has now been implicated in many human pathologies including metabolic syndrome [12,13], inflammatory bowel disease [14,15], cardiovascular disease [16] and others [17]. A more complete understanding of the mechanisms mediating host response to microbial activity within the GI tract is needed if therapeutic manipulations of the microbiota and the host responses they evoke are to be achieved.

The ancient symbiotic relationships between the host and microbial species within the GI tract are complex and human attempts to describe them use words such as mutualism, commensalism, parasitism, amensalism, and other more specialized terminology. It is becoming clear that host-microbe engagement is context dependent where environmental variables corroborate with microbial and host genetics to manage a subtle balance between the various forms of symbiotic interactions [18-21]. The history

of these interactions is, to a large extent, recorded in our genomes. These information-coding systems are both elegant and cryptic. Genomic deoxyribonucleic acid (DNA) is composed of aperiodic sequences of nucleotides that encode the instructions to develop and replicate both uni- and multi-cellular life forms. We are in the midst of a revolution in the biological sciences. The advent of “next-generation” DNA sequencing technologies [22,23] and their creative application to biological problems has profoundly altered our view of nature. High-throughput sequencing has made transformative impact in the scientific analysis of intestinal microbiota and vertebrate transcriptional genomics. However, the overlap between these two fields is fledgling at best. As DNA sequencing technologies continue to rapidly evolve, as do their democratization, the types of questions [24] open to interrogation co-evolve with them. With this perspective, I seek to highlight opportunities where advances in our understanding of the functional regions within metazoan genomes can be used to uncover the impact and history of host-microbe symbioses on human biology. I will first describe in broad detail the ontogeny of the intestinal microbiota and provide illustrative examples of the microbial contribution to nutrient metabolism and absorption in the small intestine. I then highlight how genome-wide profiling of gene expression has been an important window into the microbial impact on host physiology and pathobiology, and argue that the mechanisms underlying host transcriptional responses are poorly understood. Finally, I distill recent advances in the fields of chromatin biology and transcriptional regulation, and describe how these advances can be used to unravel microbial effects on gene regulation in the small intestinal epithelium, where I feel there is dramatic potential for knowledge gain.

In principle, my core points can be applied to the colonic epithelium as well as any cell within the animal body and in various symbiotic contexts. Inquiry into the effects of the intestinal microbiota on host biology confronts a deep and complex world of physiologic knowledge spanning microbial ecology, comparative nutrition, digestive

anatomy and physiology, and immunology, the majority of which was outside the scope of my thesis work. I touch on some of these topics and provide primary references where appropriate and comprehensive reviews where needed. The microbiota exerts a profound effect on immune system development, homeostasis, and evolution and application of transcriptional genomics to these areas will yield insight, but I have opted to focus mostly on the role of the microbiota and host transcription in the realm of nutrient metabolism where significant gaps in our knowledge exist. Throughout this Chapter and the subsequent dissertation, I attempt to highlight how viewing host-microbe symbiosis and transcriptional regulation through the lens of comparative evolution can shed light on human biology. I focus predominately on what has been learned through experimentation on mice and zebrafish as these are genetic model organisms amenable to gnotobiotics and represent approximately 450 million years of divergent co-evolution with their constituent microbiota.

2.3 The Microbiota Impacts Intestinal Physiology

2.3.1 An evolutionarily conserved developmental step

The developing vertebrate embryo is encased within the confines of a structure that creates a barrier (though not always impenetrable [25]) between self and the external environment, a fascinating feature common to nearly all metazoans. Whether a shell or womb, this structure ensures that emergence of the offspring into the outside world occurs at a defined point in development. Colonization of the vertebrate intestine by environmental microorganisms begins at the moment of emergence from this chorionic structure and a life-long interaction between host and microbe ensues. This developmental step has occurred for every generation of offspring in nature since at least the last common vertebrate ancestor (Figure 2.1).

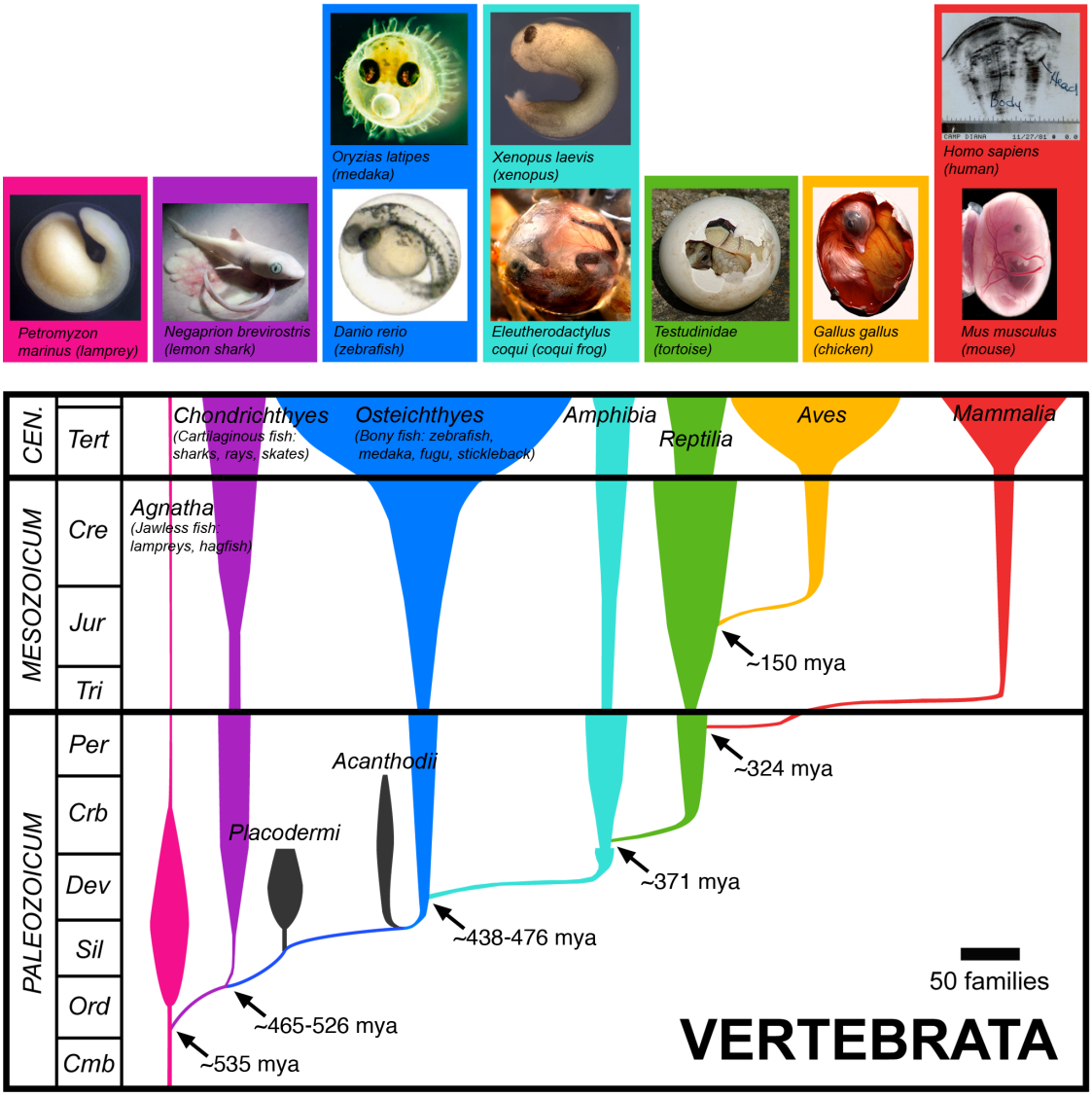


Figure 2.1: Colonization by microorganisms is an evolutionarily conserved developmental step. Vertebrates have evolved structures (chorions) that encase developing offspring in what is thought to be a predominately germ-free environment. This barrier functions in part to regulate the timing of exposure of the developing embryo to environmental microorganisms. Images show developing embryos within their protective chorions from major vertebrate classes. The spindle diagram gives rough estimates of major divergences (classes, y-axis = eras and epochs, Geologic Time Scale) and diversity (families, x-axis) based on the paleontological record. Reproduced from an open source visualization (Peter Bockman) based on [26].

2.3.2 Ontogeny of the intestinal microbiota

Microorganisms (including virus, phage, Achaea, Bacteria, protozoa, and Fungi) assemble into diverse and dynamic communities within the gastrointestinal tracts of all animals. The extent of this diversity has only been uncovered in the past decade through the application of high-throughput sequencing to gene sequences derived from small subunit ribosomal RNA (SSU rRNA; 16s rRNA in Bacteria and Achaea, and 18S rRNA in Eukarya) and using these sequences to infer phylogenetic relationships between microbes within complex communities [27-29]. This technique has spawned hundreds of studies cataloging the microbial community composition in feces and different anatomical locations of various organisms [30-32] and large-scale collaborative efforts are in progress to define the dynamics of community organization in human populations [33,34]. One of the novel outcomes has been the realization that microbial community composition (~ 160 bacterial “species” in the human gut) is governed by ecological principles derived from macro-scale ecology such as dispersal, diversification, niche construction, environmental selection, and drift [18-21,35]. These principles, along with the selective effects of host diet, lifestyle, age, and genetics, shape local microbial assemblage within the geographical context of a living multicellular host organism sensitive and responsive to its microbial habitants [9,36]. Once assembled, the intestinal microbiota and their collective genomes (microbiome [37] or metagenome [38]) function as a non-self metabolic organ dramatically affecting the metabolic potential for dietary nutrient extraction and de novo synthesis of essential nutrients.

Of particular interest is the finding that fecal microbiota from adult human monozygotic twins show no more similarity to each other than adult dizygotic twins suggesting the heritability of the microbiota, at least in humans, is low [30,39]. Fecal microbiota from biological mothers of teenage American twins showed no more similarity to their offspring than did biological fathers; furthermore, genetically un-related but co-

habitating mothers and fathers had very similar microbiotas [39]. This data suggests that the long-term effects on microbial community composition of vertical transmission of an initial inoculum of microbes from mother to offspring, as well as host genetics, are apparently not as significant as continuous environmental exposures, lifestyle, and diet in humans. Indeed, diet explains much of the variance between microbial communities when comparing across mammals [31,40] and controlled experiments have revealed diet to be a major determinate of microbial community composition [30]. For example, changing mice from a normal diet to a high fat diet results in dramatic shifts in microbial community composition within 1 day of diet change [41]. Most comprehensive studies to date have focused primarily on the low hanging fruit of the fecal microbiota from humans or mice and it should be noted that the anatomical location within the GI tract also has distinct microbial communities. Furthermore, humans and laboratory mice are anomalous in their lifestyles with respect to the rest of the natural animal kingdom whereby genetics and vertical transmission may play stronger roles in other vertebrates. Our increased knowledge of the variables controlling assembly and homeostasis of this metabolic organ illustrates a growing need to understand the functional impact of the observed microbial diversity and dynamics on host cell biology.

2.3.3 Gross anatomy of the vertebrate GI tract

The major purpose of the digestive system is to convert exogenously acquired foodstuffs into nutrients and energy required for maintenance, growth and reproduction of the animal. The source of and relative reliance on the major classes of exogenous nutrient substrates such as carbohydrates, proteins, lipids, and vitamins widely varies across animal lineages. For example, protein requirements for fish (44-60% in most *Danio rerio* diets) are much higher than those of laboratory mice (~18-20% for *Mus musculus*) [42,43]. The optimal proportion of dietary nutrients is further influenced by the

ontogeny and particular genetics of the individual organism [42,44]. Naturally, distinct evolutionary histories involving diet choice and substrate reliance have shaped the morphological and functional anatomy of vertebrate GI tracts [45]. The vertebrate GI tract has distinct functional regions along the proximal-distal axis and broad comparisons between vertebrates can best be made using nomenclature such as headgut, foregut, pancreas and biliary system, midgut, and hindgut. Compared with herbivorous ruminants, the foregut (esophagus, stomach), midgut (small intestine: duodenum, jejunum, ileum), and hindgut (cecum, proximal colon, distal colon) of omnivorous mammals such as mice and humans are anatomically similar [45]. In humans, nutrient processing and absorption has distinct anatomical hotspots along the length of the midgut. For example, carbohydrates, proteins, and lipids are absorbed in each section, but to the greatest extent in the duodenum. Conversely, bile acids and some vitamins are absorbed mostly in the ileum [46]. The zebrafish GI tract is also functionally distinct along the proximal-distal axis (foregut or anterior intestine or segment 1, midgut or middle intestine or segment 2, hindgut or posterior intestine or segment 3) however in contrast to mammals, teleost fish do not have a stomach (Figure 2.2). Similar to mammals, the majority of nutrient absorption and digestion in the zebrafish most probably occurs in the anterior intestine, whereas the zebrafish hindgut (which lacks a cecum) appears morphologically similar to the mammalian colon [47]. The extent to the functional similarity of nutrient metabolism between zebrafish and human is not entirely known and represents an interesting direction for comparative physiology. Nonetheless, it is believed that vertebrates use similar cellular pathways and molecular machines [45,47] to assimilate carbohydrates, proteins, lipids, and vitamins for use by the host.

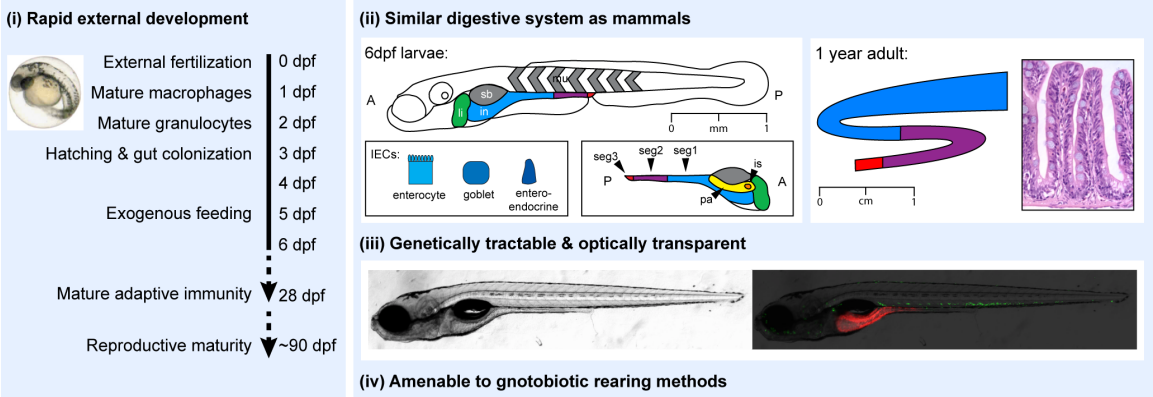


Figure 2.2: General features of the zebrafish model.
(i) Zebrafish undergo rapid and external development with a functioning GI tract at approximately 5 days post fertilization (dpf). (ii) The zebrafish digestive tract includes a liver (li), exocrine pancreas (pa), endocrine pancreatic islet (is), functionally segmented intestine (segment 1/anterior/foregut, segment 2/middle/midgut, segment 3/posterior/hindgut). Muscle (mu) and swim bladder (sb) are colored in gray. (ii) The intestinal epithelium is composed of absorptive enterocytes, goblet cells, and enteroendocrine cells. The anterior (A) and posterior (P) axes are denoted. Adult zebrafish maintain a functionally segmented intestine and have intestinal folds/villi similar to mice, but crypts are absent. (iii) Zebrafish are optically transparent and amenable to transgenesis. The image shows a 6 dpf double transgenic larvae expressing a reporter driven by an intestine specific promoter (red, *Tg(ifabp:DsRed)*) and a neutrophil specific promoter (green, *Tg(mpo:egfp)*). (iv) Zebrafish are amenable to gnotobiotic rearing techniques. At 0 dpf embryos within their protective chorions can be derived germ-free (GF) by surface sterilizing the chorion with solutions of iodine and bleach. (iv) GF animals can then be reared in sterile chambers such as cell culture flasks and fed specialized sterile diets.

2.3.4 Cellular anatomy of the vertebrate midgut

The vertebrate midgut is lined with a layer of rapidly self-renewing epithelial cells that provide the cellular interface between the host organism and the intestinal microbiota. The primary function of the midgut epithelium is to digest and absorb nutrients, provide a physical barrier against microbial infiltration to the interior of the body, and signal dietary information to other organ systems. In both fish and mammals, three major differentiated cell types (absorptive enterocytes, mucous-secreting goblet cells, and hormone-secreting enteroendocrine cells) largely perform these roles. Mice and other mammals have a fourth cell type (paneth cells) located at the base of the villi in the crypts of Lieberkühn in the small intestine, which have roles in innate immunity and secrete various bactericidal defensin peptides and lysozymes [48]. There does not

seem to be paneth cells or crypts in the zebrafish intestine [47], however bactericidal proteins such as defensins are conserved and expressed in the zebrafish intestinal epithelium [49]. Absorptive enterocytes comprise 80-90% of the midgut epithelium whereas up to 15 enteroendocrine cell subtypes are scattered throughout the mucosa comprising ~1% of epithelial cells [50]. Generally, paneth cells are enriched in the murine ileum and absent from the colon, and goblet cell number increases distally along the longitudinal axis reaching maximum numbers in the colon (~4-16%) [48]. In the zebrafish, enteroendocrine cells are observed only in the anterior intestine (segment 1), goblet cells are located in all regions, and distinct populations of enterocytes constitute the anterior versus mid/posterior intestine [47].

Terminology	Definition
Germ-free	Free from any other detectable form of life (aka. axenic)
Conventionally-raised (CONV-R)	Harboring the 'normal' indigenous, but undefined, microflora
Conventionalized (CONVD)	Ex-germfree animal colonized with a 'normal' microflora
Gnotobiotic (GN)	Describes an animal system in which all of the life forms are known
Mono, di, poly-associated (MA, DA, PA)	Ex-germfree animal harboring 1, 2, or more micro-organisms of known identity
Specific-pathogen free (SPF)	Free from pathogens, which can be specified, but otherwise with an undefined microflora

Table 2.1: Common terminology used in gnotobiotic research

2.3.5 Genetic repertoire of the intestinal microbiota

The gut microbiome (the cumulative genomes of the gut microbiota) encodes enzymes that aid in digestion of macromolecules and pathways that contribute metabolites distinct from the host repertoire [30,51-53]. In a pivotal study, Qin et al. used metagenomic sequencing to describe the microbial genes prevalent in fecal samples from 124 European individuals. This microbial gene set was 150 times larger than the

human gene counterpart and revealed a number of functional complementarities between the host genome and the microbial metagenome. For example, gut bacteria genomes are enriched for genes involved in fermentation of polysaccharides to generate energy and in the process releasing short-chain fatty acids (such as acetate, propionate, and butyrate) as by products, which are used by the host also as an energy source [3]. Microbial metagenomes are further enriched for genes involved in amino acid and lipid biosynthesis, and have capabilities for xenobiotic metabolism such as degradation of the common food supplement benzoate into pimeloyl-coenzyme-A, a precursor to biotin synthesis [52]. Intriguingly, this study also showed that only about 38% of prevalent microbial genes are common to all of the human-associated fecal microbiotas. It will be interesting to see how this trend varies along the length of the intestinal tract and in other host species. As the frontline, the intestinal epithelium senses these microbial-derived products and responds to them. Comparisons in animals with and without a microbiota are crucial to elucidate these responses.

2.3.6 Gnotobiotic vertebrate models: the power of comparison

Because evolution selected for general sterility within the chorion of the developing vertebrate embryo (Figure 2.1), researchers are able to derive animals microbe-free (germ-free, GF; Table 2.1) by sterilizing the external surface of the womb or egg and rearing the animals in an environment impervious to microorganisms. The first germ-free animal (a guinea pig) was obtained in 1895 by Nuttal and Thierfeld [54] and later entire colonies of germ-free rodents could be established [55]. Since that time germ-free derivation and rearing procedures [56] have been established for many other vertebrates including chickens, rabbits, gerbils, pigs, dogs, and zebrafish [57]. As expected, significant hurdles were encountered early on due to difficulties in maintaining general sterility and incomplete knowledge of nutrient requirements for growth and

successful reproduction. This is still an obstacle with the zebrafish and we have not yet generated sterile zebrafish colonies that can be propagated through successive generations. However, there are numerous technical difficulties associated with deriving and rearing rodents GF, and therefore requires the generation and maintenance of stable GF colonies. This provides a hurdle to using genetics to study host responses to the microbiota. In contrast, the relative ease of GF zebrafish derivation, rapid and external development, optical transparency, and genetic tractability make it an efficient system to explore host mechanisms mediating microbial responses (Figure 2.2).

Despite initial concerns about the capacity for axenic life, GF mice tend to be leaner and longer-lived than conventionally-raised (CONV-R) mice when fed appropriate diets [58]. Germ-free mice or zebrafish can be colonized with single microorganisms (mono-association, MA) or consortia of microorganisms (conventionalized, CONVD), and the microbial impact on various biological processes can be monitored. This has provided decades of phenotypic knowledge concerning the impact of the microbiota on the physiologic environment encountered by intestinal epithelial cells [3,58]. For example, the presence of a microbiota deconjugates and dehydoxylates bile acids [59], metabolizes bilirubin [60], degrades glycoproteins produced by goblet cells [61], and increases epithelial cell turnover [62]. Application of metabolomics in GF versus CONV-R/CONVD have described widespread microbial impact on the nutrient environment [53], consistent with the functional complementarities of host and microbial genomes revealed by metagenomics [52,63]. One consequence of the interplay between bacteria and host diet is an apparent increase in lipid absorption by enterocytes, which also led to increased lipid deposition in extra-intestinal tissues such as liver hepatocytes [64]. These studies highlight that the metabolic landscape encountered by intestinal epithelial cells is dynamic and dependent on microbial activities. The molecular mechanisms governing

epithelial responses to the microbiota and their evolution have not been studied with modern tools.

2.4 Transcriptional Regulation in the Intestine

An obvious yet fascinating feature that distinguishes multicellular life from unicellular counterparts is the use of a single genome to build, connect, and maintain a myriad of specialized cell types. The faithful execution of these processes requires precise spatiotemporal orchestration of gene expression. Vertebrate genomes contain billions of nucleotides (1-2 meters in length) that are assembled into linear chromosomes and packaged as chromatin into tiny spaces ($10\mu m^3$) within the cell. Protein-coding genes comprise only a small percentage (less than 5%) of nucleotides in vertebrate genomes, and the function of the remaining non-coding regions is of considerable interest [65,66]. It is becoming apparent that much of the non-coding DNA regions function to regulate cell-type specific deployment of proteins and RNAs at the level of transcription [67-70]. In this section, I review salient features of eukaryotic transcriptional regulation and discuss how sequencing based technologies are well suited for elucidating regulatory mechanisms governing host-microbe symbioses.

2.4.1 Molecular anatomy of transcriptional regulation

There has been significant progress in the field of transcriptional regulation since the discovery of RNA polymerase [71], regulatory DNA [72], and summarization of the central dogma of molecular biology [73]. A plethora of general transcription factors and mediator co-activator subunits that enable RNA polymerase assembly and recruitment to the transcription start site of genes are known [74]. Known also are checkpoints and steps governing transcription initiation, elongation, and termination [75]. We think RNA

polymerase can be paused or poised [76] and that chromatin and its remodelers play active roles throughout gene regulation [77,78]. We now observe that the genome is pervasively transcribed and that RNA can regulate its own expression [79,80]. We are even beginning to understand transcriptional events in single cells at single molecule resolution [81,82]. Of specific interest to this thesis, concerns the knowledge that specification and tuning of transcriptional activity proceeds through coordinate interactions between sequence specific transcription factors (also called transcriptional activators/repressors or transcriptional regulators) and *cis*-acting non-coding DNA (called *cis*-regulatory module, CRM or *cis*-regulatory element, CRE or *cis*-regulatory region). CRMs can be classified into two broad categories: (i) a promoter, composed of a core RNA polymerase binding sites and proximal regulatory sequences and (ii) distal regulatory regions, such as enhancers, repressors, insulators or locus control regions [83] (Figure 2.3). Both classes engage in regulatory functions, but it is the second class that appears to exhibit the majority of cell type and environment-specific regulatory control. Distal CRMs are often functionally autonomous [68] harboring binding sites for many transcription factors [84], and can be located anywhere (near the transcription start site, within introns, within exons, tens of thousands of base-pairs up or downstream, even on different chromosomes) (Figure 2.3). Chromatin looping plays a role in directly linking a distal CRM with the target gene promoter, however alternative indirect mechanisms such as place holding, spreading, and non-coding RNA intermediaries appear to better explain some experimental observations [85]. Recent work showed convincingly that intragenic CRMs could function both as an enhancer, and also as a promoter upon deletion of the proximal promoter of the target gene [86]. Identifying CRMs, discovering the transcription factors they bind, and deciphering the logic underlying their function still presents a formidable challenge within any given cell type.

2.4.2 The microbiota modulates host transcription in the intestine

Gene-specific and genome-wide profiling of gene expression has been an important window into the impact of the microbiota on host biology. In a recent comprehensive survey in the duodenum, jejunum, ileum, and colon, it was shown that the microbiota differentially regulates gene expression across the length of the GI tract (5663 cumulatively in the small intestine) [87-89]. The most highly enriched gene categories in the small intestine were associated with innate and adaptive immunity and nutrient metabolism, suggesting that many of the observed microbiota-related phenotypes have a transcriptional component. Similar results in genome-wide surveys of differentially expressed genes in GF versus CONV-R or CONVD zebrafish established a set of evolutionarily conserved transcriptional responses to the microbiota [90]. Notably, host cell response depends on the particular composition of microbes within communities in the intestine. For example, colonization of GF mice with a zebrafish microbiota elicited both overlapping and distinct transcriptome responses compared to colonization with a mouse microbiota. In fact, a microbiota transplanted from rat to mouse did not stimulate gene expression changes required for proper immune system development exemplifying the specificity of the co-evolved symbiosis between host species and their microbiota [91]. Though gene expression changes are often governed at the level of transcription, post-transcriptional regulation through alterations in mRNA splicing, stability, or translation may also be influenced by microbial activities.

It should be noted that genes do not have to be differentially regulated by the microbiota to be important for host-microbe symbiosis. This was observed for Toll-like receptor (TLR) 5, a transmembrane protein that recognizes bacterial flagellin and mediates the homeostatic balance between infection and inflammation. TLR5 was not differentially transcribed in response to the microbiota [87] yet TLR5 deficient mice have mice have an altered gut microbiota and exhibit hallmarks of metabolic disease [12]. In

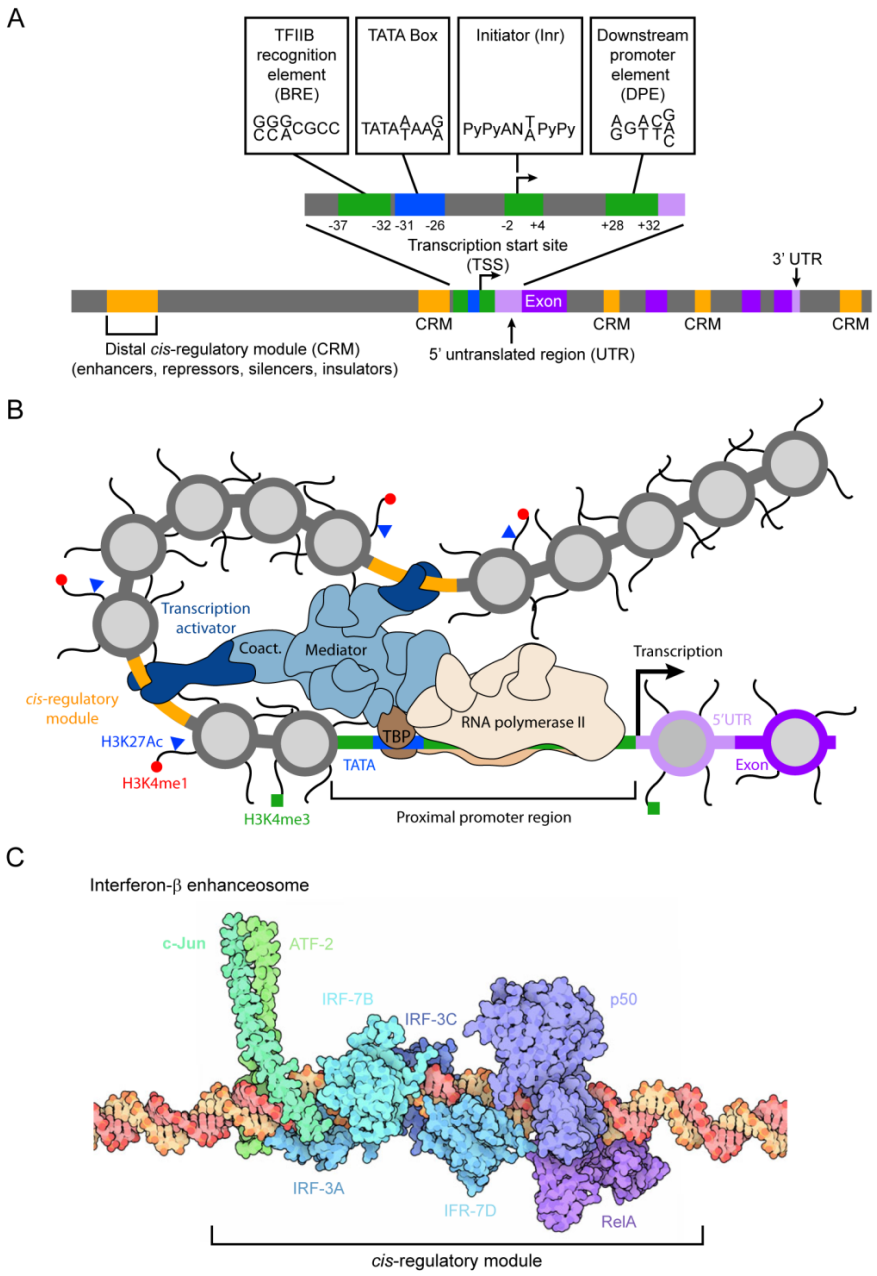


Figure 2.3: Non-coding DNA in gene regulation.
(A) Cartoon schematic of a typical gene locus illustrating types of *cis*-regulatory elements controlling gene expression. Coding exons are in dark purple, 5' and 3' untranslated regions (UTRs) are in light purple. Introns and other "non-functional" DNA are in gray. The proximal promoter is in green with core regulator regions indicated. The TATA box is in blue. Proximal and distal *cis*-regulatory modules (CRMs) are in orange. Note that CRMs can be located anywhere and function to specify spatiotemporal gene expression. (B) DNA is assembled into chromatin creating a three-dimensional layer to gene regulation. Transcription factors bind to CRMs distinguished by nucleosome depleted regions, H3K427 acetylation, and H3K4me1 methylation. Binding recruits co-activators and Mediator, which in turn create a chromatin environment conducive to RNA polymerase II binding to the proximal promoter, transcription initiation, and productive elongation. (C) Artistic visualization (reproduced with permission by David Goodsell) of the interferon-ß enhanceosome [84] revealing the complex nature and high specificity inherent to some CRMs. Here, there are 8 protein factors that make direct or indirect contact with nearly every base-pair within the 55 bp CRM.

such cases where the expressed gene mediates host-microbe symbiosis and transcription takes place without effect from the microbiota, then one might expect relatively high expression in the intestinal epithelium or associated lymphoid tissues when compared to other tissues or cell types. In this way, transcription can be regulated by mechanisms controlling cell type specificity and/or environmental response. It would be tedious to catalog the thousands of genes and their functions that are either highly expressed in the intestinal epithelium or differentially regulated by the microbiota. Instead, I highlight 10 genes that have a defined role in mediating host-microbe symbiosis (Table 2.2) on which new genomics methodologies should be able to comprehensively address their regulatory mechanics. Notably, *Angiopoietin-like 4* (*Angptl4*), a central regulator of lipid metabolism and fat storage, is suppressed specifically in the intestinal epithelium by the microbiota [8]. *Angptl4* suppression leads to increased fat storage in CONV-R and CONVD mice in comparison to GF counterparts due to its role as a direct inhibitor of Lipoprotein lipase (LPL) (Figure 2.4). This suppression is conserved in the zebrafish and therefore represents an evolutionarily ancient regulatory event. The mechanisms governing *Angptl4* intestinal transcription and microbial suppression are unknown, as is the case for the majority of genes mediating host-microbe interactions in the intestinal epithelium. In Chapter 2, I harness the unique attributes of the zebrafish (Figure 2.2) to understand the mechanisms controlling *Angptl4* suppression by the microbiota.

2.4.3 Transcription factors mediating intestinal epithelial gene expression

Sequence specific DNA-binding factors regulate gene expression at the level of transcription by selecting a gene locus for activation or repression. There are many transcription factors that are known to regulate gene expression in the intestinal epithelium. Notable examples include SMAD proteins (SMADs), caudal-related

Gene	Regulation	Tissue/Cell-type	Species	Function	REF
Angptl4	Suppressed by microbiota	Midgut (ileum)/IEC	*Dr, Mm*	Mediates microbiota-associated obesity	Bäckhed, 2004; Rawls, 2004
Alpi	Induced by microbiota	Midgut/IEC	*Dr, Mm*	Promotes mucosal tolerance to the microbiota	Cheesman, 2007
T1r3, Slglt-1, αGus	Induced by microbiota	Midgut/IEC	*Mm*	Increased sucrose intake in GF mice	Swartz, 2012
RegIIIg	Induced by microbiota	Midgut/PC	*Mm*	Bactericidal c-type lectin expressed by paneth cells	Hooper, 2006
Mucin1-4	Differentially regulated by microbiota	Midgut, Hindgut/IEC	*Mm*	Mucins maintain barrier function and are substrates for microbial symbionts	Wei, 2012; Comelli, 2007
Ang-1	Induced by microbiota	Midgut/IEC	*Mm*	Promotes microbiota induced vascular remodeling via tissue factor (TF) glycosylation	Reinhardt, 2012
α1,2-FT	Induced by microbiota	Midgut (Ileum)/IEC	*Mm*	One of many fucosyltransferases regulated by Bacteroides thetaiotaomicron	Bry, 1996
Crt1	Induced by microbiota	Midgut/IEC	*Mm*	Mediates absorption of heavy metals in intestinal epithelium perhaps through competition with microbiota	Hooper, 2002
Tlr-5	unchanged	Midgut/PC	*Mm*	Absence of expression results in microbiota-induced metabolic syndrome	Vijay-Kumar, 2010

Table 2.2: Unsolved mysteries in host-microbe symbioses

homeobox factors (Cdx), Krüpple-like factors (KLF), Hepatic nuclear factors (HNFs), Peroxisome proliferator-activated receptors (PPARs), GATA binding factors (GATA), Nuclear Factor kappa B (NFκB), Signal transducer and activator of transcription (STATs), and Suppressor of cytokine signaling (SOCS), any of which could mediate multiple aspects of host-microbe symbiosis. NFκB has been extensively studied as a central regulator of inflammatory responses to the microbiota in mammals [92,93] and zebrafish [94], and functions as an inducible transcription factor in diverse cell lineages [95]. GATA 4,5,6 are all expressed in the intestinal epithelium in mouse [96-98] and zebrafish [99,100] functioning to modulate the expression of genes involved in epithelial cell differentiation, nutrient metabolism [101] and immune responses [102]. Nuclear receptors such as PPARs [103], HNFs [104], Liver X receptor (LXR) [105,106], Vitamin

D Receptor (VDR) [107], Farnesoid X receptor/Bile acid receptor (FXR/BAR) [108], represent an important class of transcription factors that could have direct roles in host-microbe symbiosis by sensing microbial metabolites such as short-chain fatty acids and bile acid derivatives. Wnt signaling pathway transcription factors such as T cell factor (TCF) [109], Cdx1/2 [110] and Sox9 [111], and Notch signaling pathway transcription factors Hes1, Math1 [112] and Krüppel-like factors 4/5 [113] have highly conserved roles in intestinal cell-fate specification and differentiation. The Transforming growth factor-beta (TGF-ß) signaling pathway and associated transcription factors from the Smad family [114] as well as the Jak/Stat pathway and associated Stat and Socs [115] transcription factors perform integral regulatory functions in epithelial tissues to maintain mucosal integrity, renewal, and repair.

Each of these transcription factor families and associated signaling pathways are integral to intestinal epithelial cell physiology. However, most of this knowledge has been interpreted with a focus on stem cell biology, immune response, or diseases such as cancer. There is much less knowledge available concerning the role of these factors in the context of host-microbe symbiosis. As an example, searching Pubmed with the terms Wnt AND intestine gives 521 publications, whereas searching with Wnt AND microbiota gives 5 (521:5). The trend is similar for NFkB (1065:25), PPAR (391:11), TGF Beta (1249:25), Jak/stat (42:2, both are Drosophila papers), nuclear receptor (2542:34). Varying the terms or using ISI web of knowledge gives a similar story. Furthermore, searching for microbiota OR gut flora (4552) AND transcription factor gives only 136 references compared to intestine AND transcription factor (7525). Even less is known about the *cis*-regulatory modules that interpret transcription factor activity in this context. With new tools and powerful model systems in place, it is an appropriate time to fully consider the role of the commensal microbiota on intestinal transcriptional regulation.

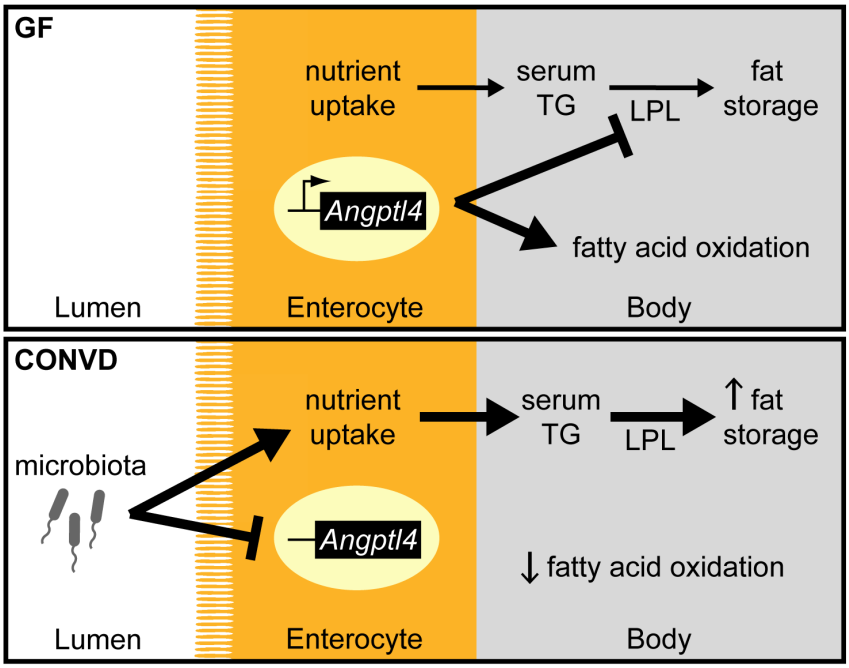


Figure 2.4: Microbial suppression of intestinal expression of *Angptl4*
Cartoon model showing that *Angiopoietin-like 4* (*Angptl4*), is suppressed specifically in the intestine of conventionalized (CONVD) mice compared to germ-free (GF) counterparts. This suppression is correlated with increased nutrient uptake in enterocytes, increased serum triacylglycerides (TG), increased lipoprotein lipase (LPL) activity, and increased fat storage in CONVD mice.

2.5 Genomic Approaches to Discover CRMs

2.5.1 Genomes as reagents for CRM discovery

Prior to the genomics age, practical discovery and functional characterization of regulatory regions was limited to DNA in close proximity to the transcription start site of single genes. The same technological advances in sequencing that have reignited interest in the intestinal microbiota have been driving the discovery of the features and functions of non-coding DNA [23,68,116]. As of writing this thesis, there are currently thirty-four publically available sequenced vertebrate genomes and soon there will be thousands more [117,118]. The diversity of sequenced vertebrates allows one to view genomes through the powerful lens of evolutionary time. Sequence alignment has revealed conservation in distal CRMs [119-121] with constraint far above what is expected for neutrally evolving DNA [122], thus implying function. Sets of conserved non-coding elements (CNEs) are enriched for known regulatory regions (enhancers, insulators, suppressors, LCRs) and many have been tested for functional enhancer activity *in vivo* [123,124]. A recent study showed that the extent to conservation (or the appearance of a conserved regulatory region in branches of the phylogeny) is dependent on the type of gene regulated [120]. This work discovered that regulatory innovations probably occurred in waves, where CNEs near genes involved in transcription factor activity or development were ancient and CNEs near genes involved in signaling pathways or post-translational protein modifications were more recent. This confirmed previous studies revealing that non-coding DNA conservation between fish and mammals is low compared to protein coding regions, and that CNEs shared between fish and mammals are enriched near developmental transcription factors [124]. There are therefore limitations to using sequence conservation as the only metric for discovering *cis*-regulatory modules. First, a conservation approach requires both distant and closely related genomes in order to distinguish signal from noise. Second, current

sequence alignment algorithms may not detect all conservation. Third, one interpretation from the Lowe et al. study could be that regulatory modules involved in certain biological processes evolve differentially. Fourth, the rules governing constraint and evolvability of *cis*-regulatory modules are not well understood. Fifth, and most importantly, conservation does not predict cell type and environment-specific activity. Therefore, other methods can be used to distinguish active non-genic DNA.

2.5.2 The role of chromatin in gene regulation

Eukaryotic genomic DNA is bundled into nucleosomes consisting of approximately 147 base pairs of DNA wrapped around 8 histone protein cores (two copies each of H2A, H2B, H3, H4) [125]. Nucleosome spacing and posttranslational modification status can have a profound effect on gene regulation [126]. Spacing proceeds through a combination of statistical positioning based on intrinsic DNA sequence affinity of the histone octamer, competition with other proteins for DNA binding, and active positioning by chromatin modifying factors [127]. Promoter regions and CRMs have the unique feature of being “open” or nucleosome depleted, allowing transcription factors and polymerases to access regulatory DNA regions. Histones are highly conserved proteins and each core histone contains a conserved unstructured tail. Histone tails can be modified post-translationally at specific residues by covalent methylation, acetylation, phosphorylation, ubiquitination, and many others [128]. A technique called chromatin immunoprecipitation (ChIP) has been applied extensively to investigate where histones with different modifications are located on DNA [129]. Briefly, in this method (I) DNA-protein complexes are cross-linked *in vivo*, (ii) cells are lysed and chromatin sheared, (ii) complexes are immunoprecipitated with an antibody targeting the protein, (iii) crosslinks are reversed, (iv) and the DNA sequence determined using PCR, microarray, or high-throughput sequencing. Recent studies have elegantly revealed that

specific histone tail residues are differentially modified dependent on the location of that nucleosome in relation to functional elements within a genome (so called histone marks) [68,130]. Importantly, these marks are also associated with the activity of the genomic region (activation, repression, pausing, poised). The histone marks H3K4me1 (active/poised) and H3K27ac (active) associate with enhancer regions and can distinguish cell type specific regulatory activity [131]. Furthermore, a number of antibodies used in mammals can function well in the zebrafish [132,133]. The same technique, ChIP-seq, can also be used to define target sites of transcription factors across the genome, though binding doesn't necessarily lead to functional output [134]. It should be noted that a new version of ChIP-seq, dubbed ChIP-exo [135], is perhaps superior to classical ChIP because it localizes the DNA binding protein to single base-pair resolution.

Two complementary genome-wide methods, DNase-seq [136,137] and FAIRE-seq [138,139], take advantage of the observation that eviction or destabilization of nucleosomes from chromatin is a characteristic feature of functional CRMs in eukaryotic genomes (Figure 2.5). DNase-seq is the genome-wide extension of the classical DNase I footprinting assay [140]. DNase I footprinting harnesses the feature that protein factors binding naked DNA block DNase I mediated enzymatic cleavage of underlying nucleotides, thus giving a quantitative footprint of the DNA binding factor. In the context of chromatin, the vast majority of DNA is protected from digestion by nucleosomes whereas regions adjacent to transcription factor binding are accessible or hypersensitive to DNase I cleavage (Figure 2.5A). This allows identification of "open" chromatin regions, which have very strong correlations with a variety of other markers (transcription factor binding, histone marks) of active non-coding regulatory function [136]. Within the "open" region defined by increased DNase I sensitivity there is often a discernible footprint of transcription factors bound to their cognate DNA sequence that is distinguished by a

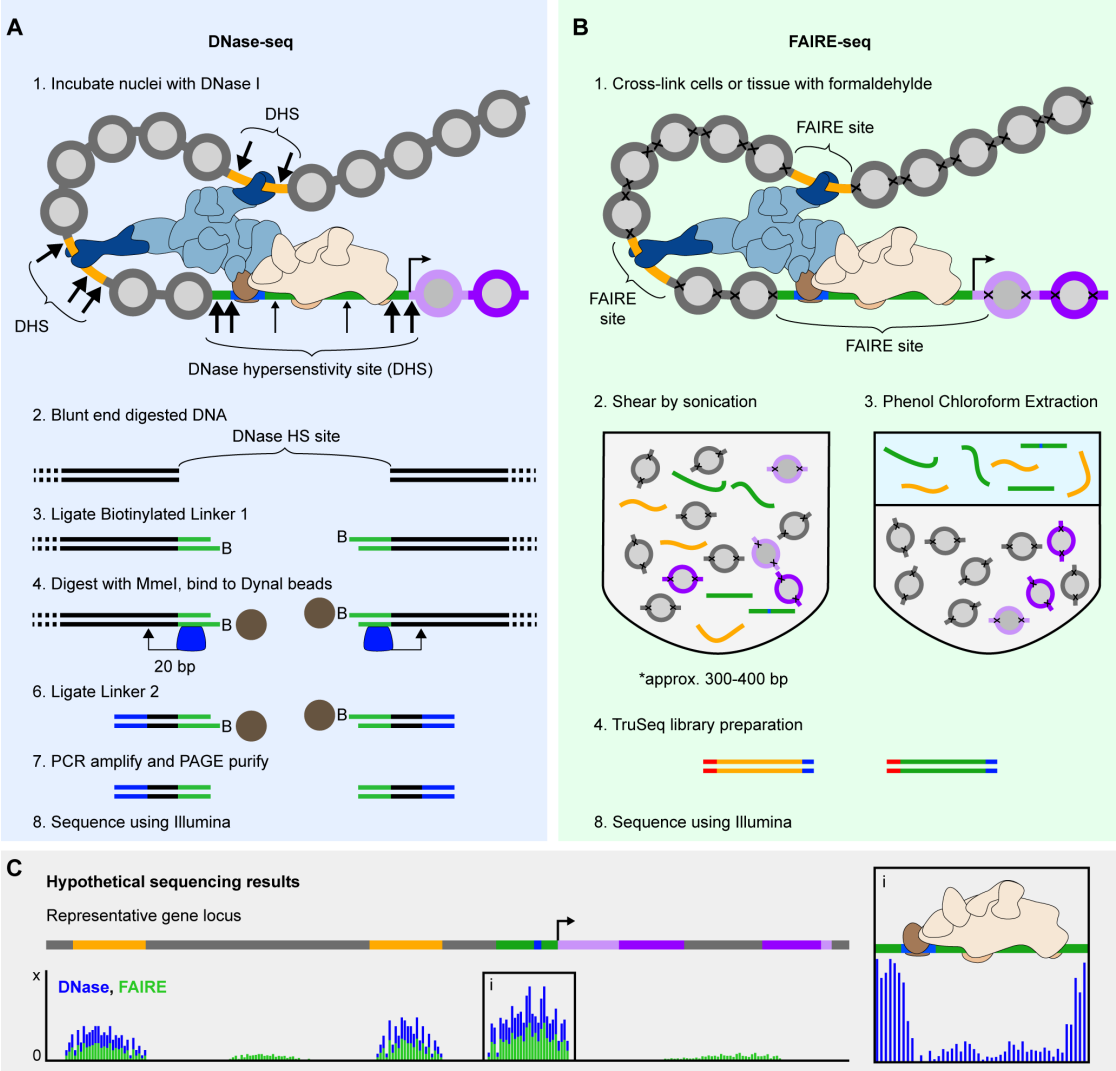


Figure 2.5: Methods for genome-wide discovery of open chromatin
(A) Flow chart describing DNase-seq methodology. Cells are lysed with weak detergent to release nuclei. Nuclei are incubated with various concentrations of DNase I endonuclease. Libraries are prepared from DNA that exhibits optimal DNase digestion (not shown). Digested DNA is blunt-ended with DNA polymerase. A biotinylated linker containing a MmeI binding site is ligated to the blunted end of digested DNA. MmeI digests DNA 20 bp away from the binding site in linker 1 and the digested product is incubated with Dynal streptavidin coated beads. Linker 2 is ligated to the blunt cut site of MmeI. The intervening DNA region is amplified by 12-16 rounds of PCR and the products are sequenced using Illumina single end reads. (B) Flow chart describing Formaldehyde Assisted Isolation of Regulatory Elements (FAIRE)-seq methodology. Cells or tissue are incubated with ~1% formaldehyde solution for a short duration (~5 minutes). Cross-linking efficiency between histones and DNA complexes is greater than that of other DNA binding factors. Cross-linked cells are lysed and chromatin sheared by sonication. Sonicated samples are phenol-chloroform extracted, which isolates free DNA (enriched for regulatory regions) into the aqueous phase and cross-linked nucleosomal DNA is trapped at the interphase. DNA fragments are prepared for Illumina single-end sequencing using the TruSeq kit. (C) Hypothetical sequencing results for a gene of interest. Exact cut sites (DNase, blue) or sequencing read counts (FAIRE-seq, green) are mapped to the appropriate reference genome sequence and visualized in a genome browser (such as UCSC). DNase-seq and FAIRE-seq peaks largely overlap [141]. DNase-seq has a high signal-to-noise ratio and distinguishes DNA-protein footprints at high resolution (inset).

local decrease in DNase I sensitivity [142]. Combined with a high signal-to-noise ratio, DNase-seq offers a powerful and validated method to discern nucleosome depleted regions as well as transcription factor-DNA interactions across the genome.

Formaldehyde-Assisted-Isolation-of-Regulatory-Elements (FAIRE) is an alternative approach to discover “open” chromatin based on differences in cross-linking efficiencies between DNA and nucleosomes compared to DNA and sequence-specific DNA-binding proteins (Figure 2.5B). In this assay, cells are covalently cross-linked briefly with formaldehyde, lysed and sonicated, and sheared chromatin is extracted with phenol/chloroform. Extraction enriches unbound DNA into the aqueous phase and protein-bound DNA is trapped to the organic/aqueous phase interface. Unbound DNA is isolated and assayed for locus-specific (via quantitative PCR) or genome-wide (via microarray, high-throughput sequencing) enrichment patterns. The signal-to-noise ratio for FAIRE-seq is not as high on average as DNase-seq, and it has yet to be proven as a method for elucidating transcription factor footprints (Figure 2.5C). However, FAIRE does not require nuclei isolation so samples do not need to be in single cell suspensions, and other experimental practicalities [139] position FAIRE-seq to be amenable to higher-throughput capabilities. Both genomics tools, DNase-seq and FAIRE-seq, can uncover a range of cell-type specific elements (promoters, enhancers, silencers, insulators, locus control regions) and do not require an antibody [141]. The impact of environmental factors, such as changes in microbial community composition and diet, on open chromatin dynamics is not well understood, and neither assay has been applied to primary intestinal epithelial cells in mouse or zebrafish.

2.5.3 Transcriptional genomics in the intestine

These studies and many others have illuminated the regulatory landscape of diverse cells in a number of pathological conditions, however chromatin genomic

datasets from the small intestinal epithelium are lacking. Recently, ChIP-seq experiments targeting GATA6, CDX2, HNF4α were performed in Caco-2 cells (a pseudo-intestinal epithelial cell line) [143,144], and a dataset for FXR in primary small intestinal epithelial cells from mouse is also available [145]. A recent study analyzed the genome-wide chromatin landscape in primary crypts and cancerous crypts from the human colon and identified enhancer activity that was lost or gained in cancerous crypts, thus explaining a majority of transcriptional changes observed in cancerous versus non-cancerous cells [146]. I suspect that similar changes to the chromatin landscape in response to microbial activity in the intestine could drive specific transcriptional programs that mediate host-microbe symbiosis. To date, there has been little effort to define the histone modification, transcription factor binding, or chromatin accessible landscape in any cell type in an axenic animal.

2.5.4 Model organisms are *in vivo* assay systems

Consortiums such as the Encyclopedia of DNA elements projects (ENCODE) and modENCODE projects have launched large-scale efforts aimed at identifying all functional elements within the genome of *H. sapiens, D. melanogaster,* and *C. elegans*. These ambitious initiatives will greatly enhance our current knowledge base for a defined set of species, tissues, and ontogenies. However, genomic regulatory responses to environmental change are fundamentally unknown, yet extremely important in all areas of biological and biomedical research. Furthermore many of the defined efforts in vertebrates will be performed in cell lines which no doubt have historic value in generating our current understanding of transcriptional regulation, but carry many caveats. It is my feeling that no existing computational, cell-culture, or *in vitro* system can fully reproduce the complexity associated with inter-cellular, inter-tissue, and inter-organ communication systems, especially in the context of host-microbe symbiosis.

Therefore, to understand *cis*-regulatory function and evolution, there is a need to perform discovery of regulatory regions in primary cells and functional characterization *in vivo* [123,147]. In this light, the intestinal epithelium has a number of features that distinguish it as an ideal genomics system to study the evolution of *cis/trans* regulatory programs mediating host-microbe symbiosis. First, the cells are abundant, accessible, and can be collected with relative ease. Second, the fundamental role of the intestinal epithelium is similar in all metazoans providing many model systems for cross-species comparisons. Third, the intestinal epithelium marks the primary interface with the microbiota and intestinal epithelial cells experience dynamic environmental factors. Fourth, despite extensive physiologic, cellular, and molecular knowledge, there is a striking paucity of information regarding genome regulation in this vital organ.

Probiotic, prebiotic, antibiotic, pharmaceuticals, strange foods and exotic beverages are consumed daily with limited knowledge of their impact on the microbial and host cells with which these foreign substance directly interface. An increased understanding of the mechanisms mediating the general impact of the microbiota on enterocyte, enteroendocrine, goblet, and paneth cell gene expression in a healthy context would pave the way for interrogation of these mechanisms in disease states. Once a set of conserved microbial response mechanisms are defined, then we can create reporter systems to determine the specific microbial signals and host signaling pathways that host cells use to perceive and respond to microbial activities. The zebrafish will be particularly useful for systematic genetic screening of microbial mutant strains [148] and chemical libraries [149,150]. A major route towards infection by pathogens is through the intestinal epithelium, and pathogens can subvert mechanisms that have evolved to maintain homeostasis between host cells and commensal microbes [151]. Furthermore, symbiosis is a sliding scale where one-time mutualists can become pathogenic if the opportunity presents itself. Therapies could be designed to shift this

balance towards mutualism if we knew better how the genetic programs have evolved to naturally maintain this balance. Genetic diseases can be caused by genic as well as non-genic mutations [152] because a mutation in a CRM may result in pathological alteration of transcript levels in the relevant tissue. Recent GWAS studies have identified numerous non-coding loci associated with intestinal disorders such as Crohn's disease and ulcerative colitis [153], yet causal mechanisms remain elusive.

Comparing functional CRMs in the intestine with other tissues would inform us to the degree of modularity in gene expression. This information could be used to design therapies that selectively alter intestinal gene expression or to predict off target effects. Transcription of genes that increase or decrease nutrient absorption in enterocytes could be targeted to combat over-nutrition or malnutrition. In a similar way, genes functioning in the biosynthesis, recognition, or secretion of hormones that are transcribed in enteroendocrine cells could be targeted to signal or inhibit satiety. It is possible that CRMs mediating microbial response are utilized in other tissues with epithelial associated microbiotas such as skin and lung and insight from the intestine could be extrapolated to these tissues [154]. The barrier and absorptive functions of trophoblasts lining the chorionic villi of the placenta are reminiscent of the intestinal epithelium and the impact of mutations, nutrients, or microbial derived products on regulatory landscape functioning in the intestinal epithelium could be predictive of seemingly disparate diseases in the placenta. Probing the impact of microbial symbiosis on intestinal gene expression in diverse fish, birds, and mammals could inform conservation, aquaculture, and agriculture rearing practices. In any case, defining a set of evolutionarily conserved host response mechanisms will maximize the utility of experimental systems, such as the gnotobiotic mouse and zebrafish, to model human health and pathology.

2.5.5 Cis-regulatory modules as mediators of host-microbe symbiosis

There is strong precedence in suggesting that changes in gene regulation can be responsible for the evolution of phenotypes. In the field of Evo-Devo (Evolution + Development), Carroll, Kingsley, Wray and many others have long argued that changes in *cis*-regulatory sequences have major roles in shaping morphological diversity [155-158]. What role does *cis*-regulatory evolution have in shaping intestinal physiology and host-microbe symbiosis? To address this question, we must profile the *cis*-regulatory genomic landscape in diverse germ-free, conventionally-raised, conventionalized, and mono-associated animals in various dietary conditions. To maximize insight, there should be systematic pipelines in place to (i) functionally test *cis*-regulatory modules for tissue-specificity and microbial response and (ii) discover associated transcription factors and signaling pathways functioning through the regulatory regions.

This was the impetus for the subsequent work presented in this thesis. In the following chapters, I first focus on a single gene involved in host-microbe symbiosis (*angptl4*) and show that the zebrafish is a powerful system for structure-function analysis of regulatory DNA *in vivo*. I then adapt methods for discovering transcription factors that function through identified *cis*-regulatory modules. Finally, I develop multiple genome-wide strategies to discover *cis*-regulatory modules in the intestinal epithelium in mouse and zebrafish. Cumulatively, this work is a major advance in our ability to probe and understand the ancient relationship between our own cells and the trillions of microbial organisms that travel through life with us.

CHAPTER 3

Intronic *Cis*-Regulatory Modules Mediate Tissue-Specific and Microbial Control of *Angptl4/Fiaf* Transcription

3.1 Overview

The intestinal microbiota enhances dietary energy harvest leading to increased fat storage in adipose tissues. This effect is caused in part by the microbial suppression of intestinal epithelial expression of a circulating inhibitor of lipoprotein lipase called Angiopoietin-like 4 (Angptl4/Fiaf). To define the *cis*-regulatory mechanisms underlying intestine-specific and microbial control of *Angptl4* transcription, we utilized the zebrafish system in which host regulatory DNA can be rapidly analyzed in a live, transparent, and gnotobiotic vertebrate. We found that zebrafish *angptl4* is transcribed in multiple tissues including the liver, pancreatic islet, and intestinal epithelium, which is similar to its mammalian homologs. Zebrafish *angptl4* is also specifically suppressed in the intestinal epithelium upon colonization with a microbiota. *In vivo* transgenic reporter assays identified discrete tissue-specific regulatory modules within *angptl4* intron 3 sufficient to drive expression in the liver, pancreatic islet β-cells, or intestinal enterocytes. Comparative sequence analyses and heterologous functional assays of *angptl4* intron 3 sequences from 12 teleost fish species revealed differential evolution of the islet and intestinal regulatory modules. High-resolution functional mapping and site-directed mutagenesis defined the minimal set of regulatory sequences required for intestinal activity. Strikingly, the microbiota suppressed the transcriptional activity of the intestine-specific regulatory module similar to the endogenous *angptl4* gene. These results

suggest that the microbiota might regulate host intestinal Angptl4 protein expression and peripheral fat storage by suppressing the activity of an intestine-specific transcriptional enhancer. This study provides a useful paradigm for understanding how microbial signals interact with tissue-specific regulatory networks to control the activity and evolution of host gene transcription.

3.2 Introduction

The vertebrate intestine harbors a dense community of microorganisms (gut microbiota) that exerts a profound influence on distinct aspects of host physiology [20,37]. The gut microbiota has been identified as a potent environmental factor in a growing number of human diseases, including inflammatory bowel disease [15], antibiotic-associated diarrheas [16], cardiovascular disease [16], and obesity [9]. As a consequence, there is considerable interest in understanding the mechanisms by which this resident microbial community influences health and disease in humans and other animals.

The ability of the microbiota to modify host nutrient metabolism and energy balance is a prominent theme in host-microbe commensalism in the intestine. Recent mechanistic insights into this process have been provided by comparisons between mice reared in the absence of microbes (germ-free or GF) to those colonized with members of the normal microbiota, as well as high-throughput DNA sequencing analysis of the metabolic potential of gut microbial genomes. These approaches have shown that the gut microbiota contributes biochemical activities not encoded in the host genome that enhance digestion of dietary nutrients [52,159]. The resulting increase in digestive efficiency results in elevated plasma levels of triglyceride (TG)-rich lipoproteins [8,11]. TG within circulating lipoprotein particles is hydrolyzed through the rate-limiting activity of lipoprotein lipase (LPL) located at the luminal surface of capillaries. TG hydrolysis

releases free fatty acids (FFA) for uptake by adjacent tissues for oxidation (e.g., in cardiac and skeletal muscle) or fat storage (e.g., in adipose tissues) [160]. The presence of a gut microbiota also results in a concomitant reduction in intestinal expression of *Angiopoietin-like 4* (*Angptl4*, also called *Fiaf*, *Pgar*, and *Hfarp*) [8,161], encoding a circulating peptide hormone that acts as a direct inhibitor of LPL activity [162-165]. Studies in gnotobiotic mice have indicated that microbial suppression of *Angptl4* expression is restricted to the intestinal epithelium and is not observed in other tissues that express *Angptl4,* such as liver and adipose tissue. This restricted suppression leads to a significant increase in LPL activity and fat storage in adipose tissue of animals colonized with a microbiota, which is an effect abolished in mice lacking *Angptl4* [8]. These results have established *Angptl4* as a key host factor mediating the microbial regulation of host energy balance and have raised considerable interest in defining the mechanisms underlying the tissue-specific and microbial regulation of *Angptl4* expression. The importance of understanding mechanisms regulating Angptl4 production is further underscored by reports suggesting that human ANGPTL4 functions as an important determinant of plasma TG levels [166,167] and by Angptl4's additional functions in angiogenesis [168], tumor cell survival [169] and metastasis [170,171], and wound healing [172].

Previous studies have revealed that mammalian *Angptl4* expression is subject to complex cell type-specific regulation but the underlying mechanisms remain unclear. *Angptl4* mRNA in humans and rodents is expressed in multiple tissues, including adipose tissue, liver, intestinal epithelium, pancreatic islets, and cardiac and skeletal muscle [8,169,173-176]. Preliminary insights into the *trans*- and *cis*-regulatory mechanisms controlling *Angptl4* transcription have been provided by analyses in non-intestinal tissues. Members of the peroxisome proliferator-activated receptor (PPAR) family of nuclear receptors (i.e., PPARγ, PPARα, and PPARβ/δ) have been identified as

activators of *Angptl4* expression in adipose tissue, liver [173,177], skeletal [178] and cardiac muscle [179], myofibroblasts [180], and colon carcinoma cells [181]. A PPAR-responsive element (element defined as a transcription factor binding site or TFBS) located in the proximal portion of *Angptl4* intron 3 has been shown to directly bind different PPAR family members in adipose tissue, liver [177], and myofibroblasts [180]. Additional studies in non-intestinal cell types have identified functional TFBSs for SMAD3 and glucocorticoid receptor in the 5' distal region and 3' untranslated region (UTR), respectively [180,182]. *Angptl4* transcription is induced under hypoxic conditions in several non-intestinal cell types by hypoxia-inducible factor 1α (HIF1α) [183,184]; however, the TFBSs mediating this response have not been identified. These studies support a role for these *trans*- and *cis*-regulatory factors in controlling *Angptl4* transcription in these cell types, yet the mechanisms underlying the transcription of *Angptl4* in other tissues, such as the intestine and pancreatic islet, remain unknown. Moreover, the *cis*/*trans*-regulatory mechanisms underlying microbial suppression of *Angptl4* transcription in the intestinal epithelium remain undefined.

The zebrafish (*Danio rerio*) provides unique opportunities to study the transcriptional regulatory programs mediating tissue-specific and the microbial control of vertebrate gene expression. Robust transgenesis methods using the Tol2 transposon system [185], large numbers of offspring, and optical transparency facilitate efficient spatiotemporal analysis of reporters driven by potential DNA regulatory regions in mosaic and stable transgenic animals [186]. The anatomy and physiology of the zebrafish digestive tract are highly similar to mammals, including an intestine, liver, gall bladder, and exocrine and endocrine pancreas [187-189]. The intestinal epithelium of the zebrafish displays proximal-distal functional specification and is composed of absorptive enterocytes as well as secretory goblet and enteroendocrine lineages [47,190]. The zebrafish intestine is colonized by a microbiota shortly after the animals hatch from their

protective chorions at 3 days post-fertilization (dpf) [191,192] and reaches a stage sufficient to support nutrient digestion by 5 dpf [193]. To study the roles of commensal microbes on zebrafish development and physiology, we have developed methods for rearing GF zebrafish and colonizing them with members of the normal zebrafish microbiota [57,194]. By combining these methods with functional genomic approaches, we identified zebrafish transcripts that display altered expression levels in animals raised GF compared to those colonized with a normal microbiota, including microbial suppression of a zebrafish homolog of mammalian *Angptl4* [90,94,195]. The expression pattern of this zebrafish *Angptl4* homolog, and the mechanisms underlying the tissue-specific and microbial regulation of its expression, have not been previously described.

These features position the zebrafish as a powerful model for assaying the regulatory potential of DNA involved in mediating cell-specific and microbe-responsive transcriptional events. Previous studies of DNA regulatory potential in the zebrafish system have focused primarily on developmental genes [196-200], and it remains unclear if the lessons learned from these analyses [201] will apply to physiologic genes like *Angptl4* that are regulated by endogenous as well as exogenous cues. Moreover, a paucity of available genome sequences for teleost species closely related to zebrafish has severely limited prior evolutionary analysis of *cis*-regulatory sequence and function. Here, we utilize the zebrafish to investigate the *cis*-regulatory mechanisms governing tissue-specific and microbial control of *Angptl4* transcription. We focus our analysis on intestinal and islet expression, where the mechanisms regulating *Angptl4* transcription have not been adequately examined. We first uncover distinct intronic *cis*-regulatory modules (CRM, defined here as a discrete DNA region containing sufficient information to confer a regulatory function) that mediate intestinal and islet expression. Using this information, we reveal that the intestine-specific CRM also responds to microbial stimuli to suppress *angptl4* expression. These results provide novel insights into how

vertebrates might control the tissue-specific transcription of *Angptl4* and constitute an important advance towards understanding how commensal gut microbes regulate gene expression and energy balance in their vertebrate hosts.

3.3 Results

3.3.1 Tissue-specific expression of zebrafish *angptl4*

A comparative sequence analysis revealed that the zebrafish genome encodes a single ortholog of mammalian Angptl4 that displays marked amino acid sequence conservation with other vertebrate homologs (Figures 3.1A, 3.S1, 3.S2). We used RNA whole-mount *in situ* hybridization (WISH) to identify the tissues in which *angptl4* is transcribed during zebrafish development. We found that zebrafish *angptl4* mRNA is expressed ubiquitously in 1 dpf embryos (Figure 3.1B) but becomes enriched in specific tissues during post-embryonic stages. Transcripts for *angptl4* are enriched in the intestinal epithelium by 4 dpf, shortly after the intestinal tract becomes completely patent (Figure 3.1C), and become localized to the anterior intestine (segment 1) by 6 dpf (Figure 3.1D,E). Transcripts for *angptl4* were also enriched in the pancreatic islet by 8 dpf (Figure 3.1F) and in the liver by 17 dpf (Figure 3.1G,H). Notably, the intestinal epithelium [8,161], liver [174,177], and pancreatic islet [175] in mammals also express *Angptl4* mRNA. These data establish that the zebrafish *angptl4* ortholog is expressed in a tissue-specific pattern that is conserved across vertebrate lineages and suggest that the underlying transcriptional regulatory mechanisms may also be conserved.

3.3.2 Conservation in DNA sequence guides *cis*-regulatory module discovery

Previous studies have indicated that conservation in non-coding genomic DNA sequence across vertebrate lineages can be a reliable predictor of *cis*-regulatory DNA regions [202,203]. We therefore used this approach to discover regulatory regions

controlling transcription of *angptl4* in the liver, islet, and intestinal epithelium. Mammals and teleost fishes diverged approximately 438-476 million years ago [204], whereas zebrafish (clade Otocephala) diverged from other teleost fishes with currently-available genome sequence {clade Euteleostei; i.e., medaka (*Oryzias latipes*), stickleback (*Gasterosteus aculeatus*), fugu (*Takifugu rubripes*), and tetraodon (*Tetraodon nigroviridis*)} approximately 230-307 million years ago [205]. We generated multiple-species LAGAN alignments with Vista software using 10 kb of genomic sequence surrounding and including the *angptl4* loci from four teleost fishes (zebrafish, medaka, tetraodon, fugu) and three mammals {human (*Homo sapiens*), dog (*Canis familiaris*), and mouse (*Mus musculus*)}. Alignment of teleost and mammalian genomic sequences did not detect regions of primary sequence conservation within *angptl4* non-coding regions (>50% over 100 bp; data not shown), suggesting that these alignment methods are not sufficiently sensitive to detect existing non-coding conservation [202] or that the composition and/or location of non-coding regulatory regions are not stringently conserved between these lineages. We therefore separately aligned teleost *angptl4* (Figure 3.2A) and mammalian *Angptl4* loci (Figure 3.2B) and searched for non-coding sequence conservation in each lineage. These alignments revealed that human and zebrafish *angptl4* loci both contain 7 conserved exons as well as a concentration of conserved non-coding sequences directly upstream of exon 1 and in intron 3 **(**Figure 3.2). Similarities in gene structure and locations of conserved non-coding regions, in addition to conservation in gene expression patterns, support the hypothesis that the regulatory mechanisms of *angptl4* transcription may be evolutionarily conserved.

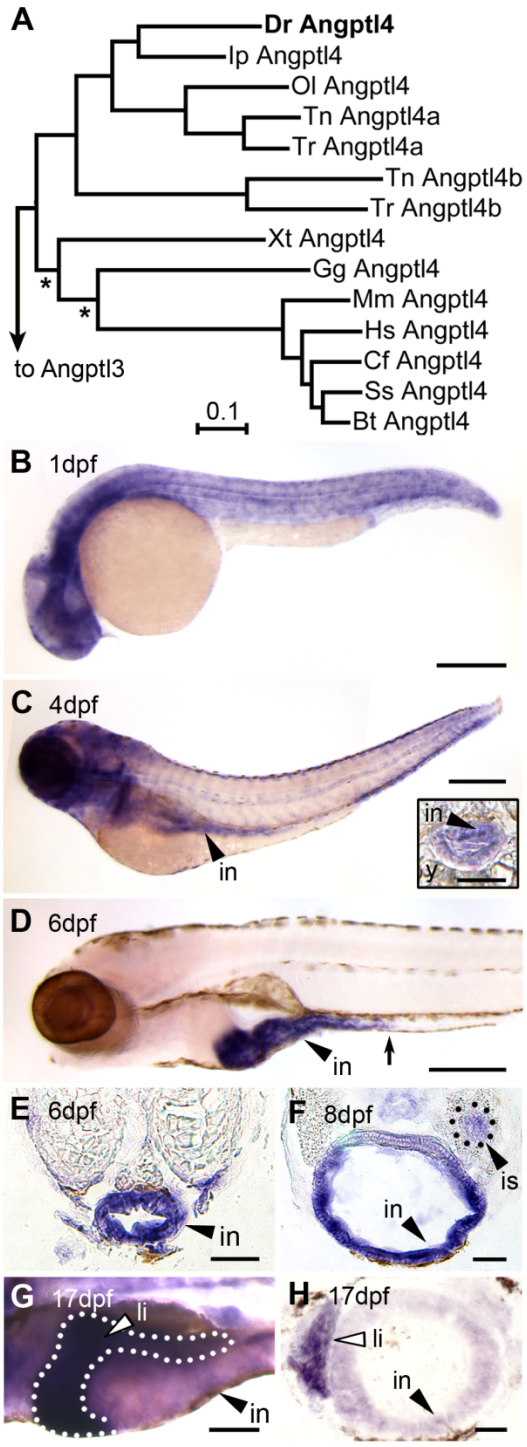


Figure 3.1: Tissue-specific expression of zebrafish *angptl4* mRNA.
(A) Distance phylogram of Angptl4 protein from zebrafish *(Dr*, *Danio rerio),* catfish (*Ip*, *Ictalurus punctatus*), medaka *(Ol*, *Oryzias latipes),* tetraodon (*Tn*, *Tetraodoan nigroviridis),* fugu (*Tr*, *Takifugu rubipres*), xenopus (*Xt*, *Xenopus tropicalis*), chicken (*Gg*, *Gallus gallus*), mouse (*Mm*, Mus *musculus*), human (*Hs*, *Homo sapiens*), dog (*Cf*, *Canis familiaris*), pig (*Ss*, *Sus scrofa*), cow (*Bt*, *Bos taurus*). All nodes are significant (>700/1000 bootstrap replicates) except those marked with an asterisk (*). Scale bar indicates phylogenetic distance, in number of amino acid substitutions per site. We found that the genomes of zebrafish, channel

catfish (*Ictaluris punctatus*), and medaka (*Oryzias latipes*) encode a single ortholog of mammalian Angptl4, whereas two pufferfish species (*Takifugu rubripes* and *Tetraodon nigroviridis*) encode two Angptl4 paralogs. See also Figure 3.S1. (B-G) Whole-mount *in situ* hybridization (WISH) using a riboprobe targeting *angptl4* mRNA during various stages in zebrafish development reveals dynamic spatiotemporal gene expression patterns. (B) At 1 day post fertilization (dpf) embryos exhibit ubiquitous expression of *angptl4*. (C-D) By 4 dpf, marked expression is observed in the intestinal epithelium (in, black arrowhead), but by 6 dpf, robust expression becomes largely localized to the intestine (black arrowhead) and pancreatic islet (not shown). The black arrow marks the boundary between the anterior intestine (segment 1) and mid-intestine (segment 2). Scale bars = 500 μm. (E-F) Transverse sections of 6 dpf and 8 dpf animals confirm expression in the intestinal epithelium (E, in, black arrowhead) and pancreatic islet (F, is, black triangle). Scale bars = 50μm. (G-H) At 17 dpf, strong expression is observed in the liver (li, white arrowhead, dotted line outlines the liver). G, Scale bar = 250 μm; H, Scale bar = 50 μm.

3.3.3 The *angptl4* proximal promoter does not recapitulate mRNA expression patterns

We assayed the regulatory potential of DNA upstream and proximal to the zebrafish *angptl4* transcription start site (TSS) for the ability to transcribe a reporter in the intestine, liver, and islet. We first employed 5' rapid amplification of cDNA ends (5'RACE) to determine the location of the TSS (Figure 3.S3B). We identified a single TSS located 89 base pairs (bp) upstream of the translation start site and a canonical TATA box at position -31 bp of the TSS (Figure 3.S3B). Based on this analysis and expressed sequence tag (EST) coverage of the zebrafish *angptl4* locus (data not shown), we found no evidence of alternative promoters farther upstream of the defined TSS. Using Tol2 transposon transgenesis, we assayed the regulatory potential of genomic DNA upstream of the zebrafish *angptl4* TSS, including the 5' untranslated region (UTR) (Figure 3.S3A), to drive expression of an enhanced green fluorescent protein (GFP) reporter in 0-7 dpf zebrafish larvae. We found that regulatory DNA within -1 kb, -3.5 kb, or -5.2 kb upstream of the TSS harbors the potential to drive GFP expression in mosaic animals in several tissues including liver at 6 dpf (Figure 3.S3C,E). Robust expression in the liver was confirmed in animals harboring stable germ-line incorporation of these transgenes (Figure 3.S3D,F). However, these *angptl4* upstream

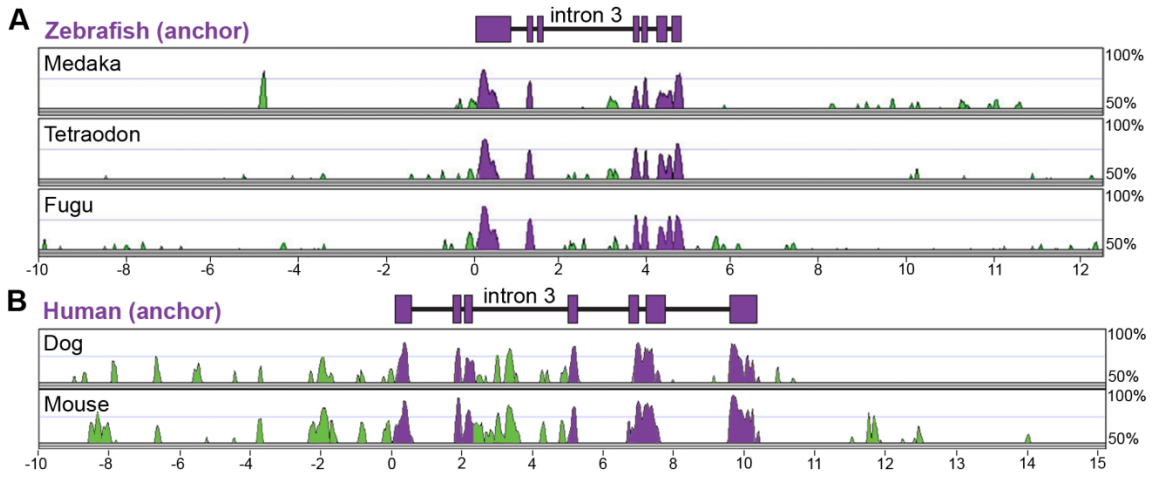


Figure 3.2: Multiple-species alignments reveal conservation in *angptl4* gene structure and location of conserved non-coding regions.
(A) VISTA plot displaying the global pairwise alignment of the zebrafish *angptl4* locus with the orthologous medaka, tetraodon, and fugu regions and (B) human *ANGPTL4* locus with the orthologous mouse and dog regions. Purple conservation peaks correspond to exonic sequences, and green conservation peaks represent non-coding sequences. The zebrafish and human gene structure are denoted by purple boxes above the corresponding VISTA plot (VISTA parameters: 100 bp sliding window, LAGAN alignment). Note that the concentration of conservation peaks within intron 3 of both teleost and mammalian *angptl4* genes.

regulatory sequences were not sufficient to drive detectable reporter expression in the intestine (Figure 3.S3G) or islet (data not shown). We therefore reasoned that information governing transcription in the intestine and islet must be located distal to the TSS and proximal promoter.

3.3.4 *Angptl4* intronic CRMs confer tissue-specific transcription

Relatively high levels of DNA sequence conservation in both teleost and mammalian lineages (Figure 3.2) prompted us to test the 3rd intron of zebrafish *angptl4* for transcriptional regulatory potential. We cloned full-length zebrafish *angptl4* intron 3 (2,136 bp; designated in3) into a Tol2 transposon reporter vector upstream of a minimal mouse *Fos* promoter (*Mmu.Fos*) driving transcription of a GFP or tdTomato reporter. Importantly, the minimal *Fos* promoter alone is relatively inactive in most tissues and is

not sufficient to drive transcription of detectable levels of GFP in the intestine, islet, or liver [186]. Analysis of 6 dpf zebrafish larvae with mosaic expression of the *Tg(in3-Mmu.Fos:GFP)* transgene disclosed that full-length in3 is sufficient to confer reporter expression in multiple tissues including the liver, muscle, intestine (Figure 3.3C), and islet (not shown). This expression pattern was confirmed in fish with stable germ-line incorporation of the transgene (Figure 3.3D). Guided by sequence conservation between zebrafish and medaka (Figure 3.2A), we assayed serial truncations of in3 for spatial regulatory potential to determine whether reporter transcriptional activity in these distinct tissues is governed by the same CRM or through multiple discrete CRMs, (Figure 3.3B). The first truncation separated liver expression (1,219 bp, designated in3.1, Figure 3.3E,F) from islet and intestinal expression (701 bp; designated in3.2, Figure 3.3G,H). Further truncation of in3.2 uncoupled islet (387 bp; designated in3.3; Figure 3.3I,J) and intestinal (316 bp; designated in3.4; Figure 3.3K,L) expression. This analysis therefore revealed non-overlapping modules sufficient to confer mosaic and stable reporter expression in the liver, islet, and intestinal epithelium that is consistent with endogenous *angptl4* mRNA expression (Figure 3.1).

We next sought to identify the specific cell types in the intestinal epithelium and pancreatic islet in which modules in3.3 and in3.4 respectively enhance transcription. To define the cell type within the islet in which module in3.3 is active, we utilized a zebrafish transgenic line that drives expression of cyan fluorescent reporter (CFP) specifically in insulin-producing β-cells within the islet (*Tg(ins:CFP-NTR)*^{s892}) [206]. *In vivo* imaging of 6 dpf progeny from intercrosses of *Tg(ins:CFP-NTR)*^{s892} and *Tg(in3.2-Mmu.Fos:tdTomato)* adults revealed strong co-localization of CFP and tdTomato (Figure 3.3O), indicating that the in3.3 module specifically enhances transcription in pancreatic β-cells. Immunofluorescence assays of sectioned 6 dpf zebrafish stably expressing the *Tg(in3.4-*

Mmu.Fos:GFP) transgene revealed that GFP driven by the in3.4 module co-localizes with 4E8-positive absorptive enterocytes (Figure 3.3M) but not with 2F11-positive secretory cells in the intestinal epithelium (Figure 3.3N). These data suggest that in3.4 functions as an enterocyte-specific transcriptional regulatory module.

We next tested whether the intestine-specific reporter expression generated by module in3.4 is independent of the *Fos* minimal promoter, orientation, and proximal position to the TSS. This module is located downstream of the TSS in intron 3 of the endogenous *angptl4* gene; however, our synthetic reporter construct positions it upstream of the TSS and the *Fos* minimal promoter. We therefore cloned in3.4 into a position downstream of *GFP* in either the forward or inverse orientation under control of either a *Fos* minimal promoter or the -1 kb *angptl4* promoter. Each of these constructs was sufficient to promote robust reporter expression in the anterior intestine of 6 dpf mosaic and stable zebrafish (Figure 3.S4A and data not shown), similar to our observations with in3.4 located in the proximal position (Figure 3.3K,L). These results establish that in3.4 is a bona fide transcriptional enhancer module active in enterocytes in the anterior intestine. We next used DNase I hypersensitivity to determine if the in3.4 module functions as an intestinal regulatory module *in vivo* at the endogenous *angptl4* locus. To obtain a sufficient number of intestinal epithelial cells for this assay, we analyzed intestines from adult zebrafish. Stable transgenic zebrafish harboring the in3.2 or in3.4 reporter maintain reporter activity in the intestine into adulthood (Figure 3.S4B and data not shown) indicating this module and associated *trans*-regulators are active in the adult zebrafish intestine. We find that the endogenous *angptl4* promoter and in3.4 module, but not the adjacent in3.3 module, are hypersensitive to DNase I cleavage in intestinal epithelial cells isolated from adult zebrafish (Figure 3.3P). The endogenous in3.4 module is therefore an active regulatory module in the intestinal epithelium, under

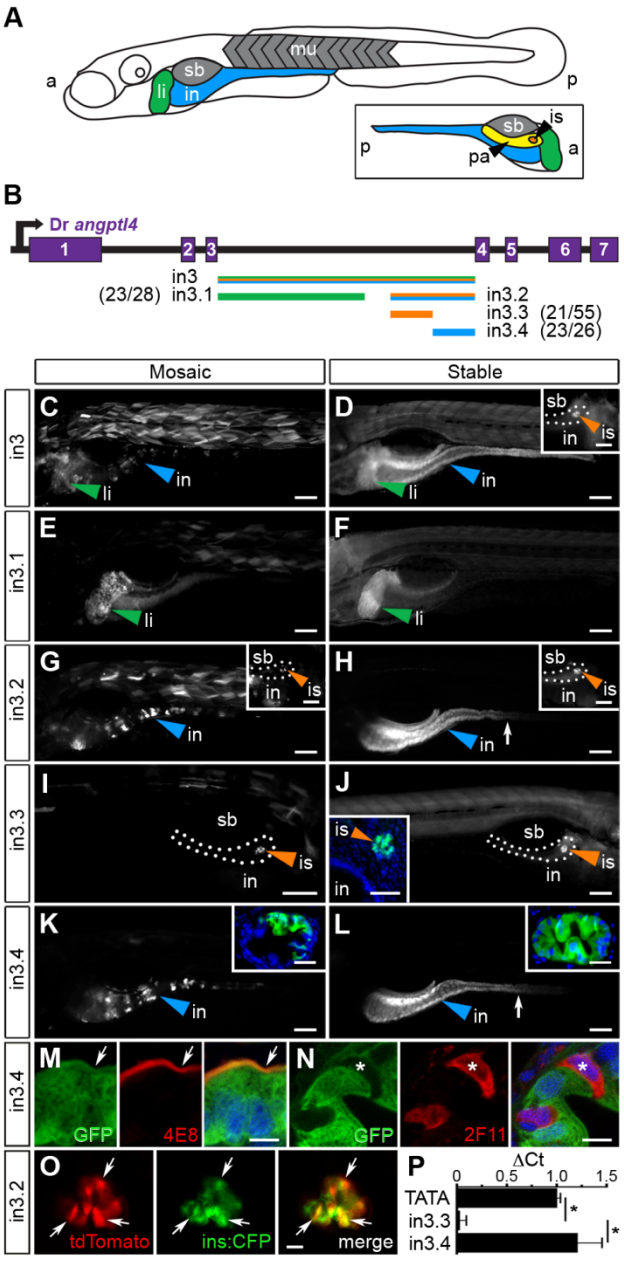


Figure 3.3: Non-overlapping regulatory modules within *angptl4* intron 3 confer liver, islet, and enterocyte-specific reporter expression.

(A) Depiction of the 6 dpf zebrafish showing liver (li, green), intestine (in, blue), swim bladder (sb, grey), and muscle (mu, grey), with the fish oriented anterior (a) to the left and posterior (p) to the right. The opposite orientation reveals the exocrine pancreas (pa, yellow) and islet (is, orange). (B) Scaled schematic of the zebrafish *angptl4* locus and non-coding DNA assayed for regulatory potential. Modules are color coded according to the tissues in which they confer expression. Ratios of islet or intestine positive fish versus total fish expressing gfp are shown in parentheses next to truncation labels. (C-N) Representative images of GFP reporter expression in mosaic (column 1) and F_1 stable (column 2) animals driven by each non-coding DNA region (rows). Scale bars = 100 μm; li = liver, is = islet, in = intestine, sb = swim bladder. Colored arrowheads indicate tissue with specific reporter expression. (C-D) Full-length intron 3 (in3; 2,136 bp) is sufficient to promote expression of the reporter in the liver, islet (D, inset, scale bar = 50 μm), and intestine. (E-F) Truncation in3.1 (1,219 bp) confers expression in the liver. (G-H) Truncation in3.2 (701 bp) confers

expression in both the intestine and islet (H, inset). Inset scale bar = 50 μm. (I-J) Truncation in3.3 (387 bp) confers islet expression. A transverse section (inset, J) reveals islet expression (nuclei stained with DAPI). Inset scale bar = 50 μm. (K-L) Truncation in3.4 (316 bp) confers intestinal expression. Insets in panels K and L contain transverse sections showing expression localized to the intestinal epithelium (nuclei stained with DAPI). Inset scale bar = 25 μm. The dotted lines in panels D, G, H, and I outline the pancreas. The white arrows in panels H, K, and L mark the boundary between the anterior intestine (segment 1) and mid-intestine (segment 2). (M-N) Cells expressing GFP driven by the in3.4 regulatory module colocalize with a marker (4E8, red, white arrow) of the brush border of absorptive enterocytes, but fail to co-localize with marker for secretory cells (2F11, red, asterisk). Nuclei stained with DAPI. Scale bars = 5 μm. (O) Intercross of *Tg(in3.2-Mmu.Fos:tdTomato)* with β-cell specific reporter line (*Tg(ins:CFP-NTR)*^{s892}) show colocalization of tdTomato and CFP in the islet. Scale bars = 10 μm. (P) Quantitative PCR shows that the in3.4 module and the *angptl4* promoter (TATA box), but not the in3.3 module, are hypersensitive to DNase I cleavage in intestinal epithelial cells isolated from adult zebrafish. Asterisks denote P-value <.01 from unpaired T-tests between TATA box or in3.4 and in3.3 regions. Error bars represent standard deviation from four biological replicates using cells pooled from 3 wild-type adult zebrafish per replicate.

regulatory control distinct from the adjacent in3.3 module, consistent with our transgenic reporter analysis of this same region. Together, these data reveal extensive transcriptional regulatory potential within intron 3 of zebrafish *angptl4* and suggest that distinct intronic modules may mediate spatially restricted transcription of *angptl4* in the intestinal epithelium, pancreatic β-cells, and liver.

3.3.5 Evolution of the islet and intestinal regulatory modules

We used comparative genome sequence analysis from 12 teleost fishes and heterologous *in vivo* reporter assays to explore the evolution of the islet and intestinal regulatory modules. We originally postulated that evolutionary conservation of non-coding sequences could be used to predict the location of *cis*-regulatory regions controlling spatial and environmental regulation of *angptl4* transcription (Figure 3.2). However, the significant amount of time (approximately 230-307 million years ago) [205] since the divergence between zebrafish (clade Otocephala; order Cypriniformes) and the other teleost fish with available genome sequence (all from clade Clupeocephala, such as medaka) did not permit high-resolution analysis of recent evolution of zebrafish *angptl4* regulatory sequences (Figure 3.2A). We therefore sequenced the intronic region

orthologous to in3.2 from 10 additional Ostariophysi species, including 1 from order Siluriformes (channel catfish, *Ictalurus punctatus*) and 9 other members of order Cypriniformes (Figure 3.4A). Because genome sequences are not currently available for these species, we took advantage of the intronic location of these regulatory modules by utilizing PCR primers targeting highly conserved sequences in flanking exons 3/4 or intron 3 to clone and sequence these putative regulatory regions. As expected, pairwise alignments of new sequences orthologous to zebrafish in3.2 revealed an inverse relationship between the phylogenetic distance between the two species and module sequence conservation, with the intestinal module diverging more rapidly than the islet module (Figures 3.4B, 3.S5, 3.S6). To test the functional consequences of the observed module divergence in these teleost species, we analyzed each module using our zebrafish mosaic transgenic assay for regulatory potential in the intestine and islet. Despite accounts of functional conservation in the absence of primary sequence conservation [196,207], the non-coding sequence within medaka *angptl4* intron 3 orthologous to zebrafish in3.2 (Ol in3.2) failed to drive reporter expression in either the reporter expression in the islet (Figure 3.4C). However, only in3.2 from Cypriniformes intestine or islet (Figure 3.4C). Notably, all tested Ostariophysi modules elicited robust species within the Danio monophyletic group (*Danio nigrofasciatus, D. choprae, D. feegradei*) [208,209] were sufficient to confer reporter expression in the intestine (Figure 3.4C) despite marked regions of sequence conservation within the intestinal module in other Cypriniformes species (*D. aequipinnatus*, *C. auratus*, *C. carpio*, *P. conchonius*). These results reveal differential evolutionary dynamics of the *angptl4* intestinal and islet modules and support the hypothesis that high sequence conservation is required for tissue-specific transcription.

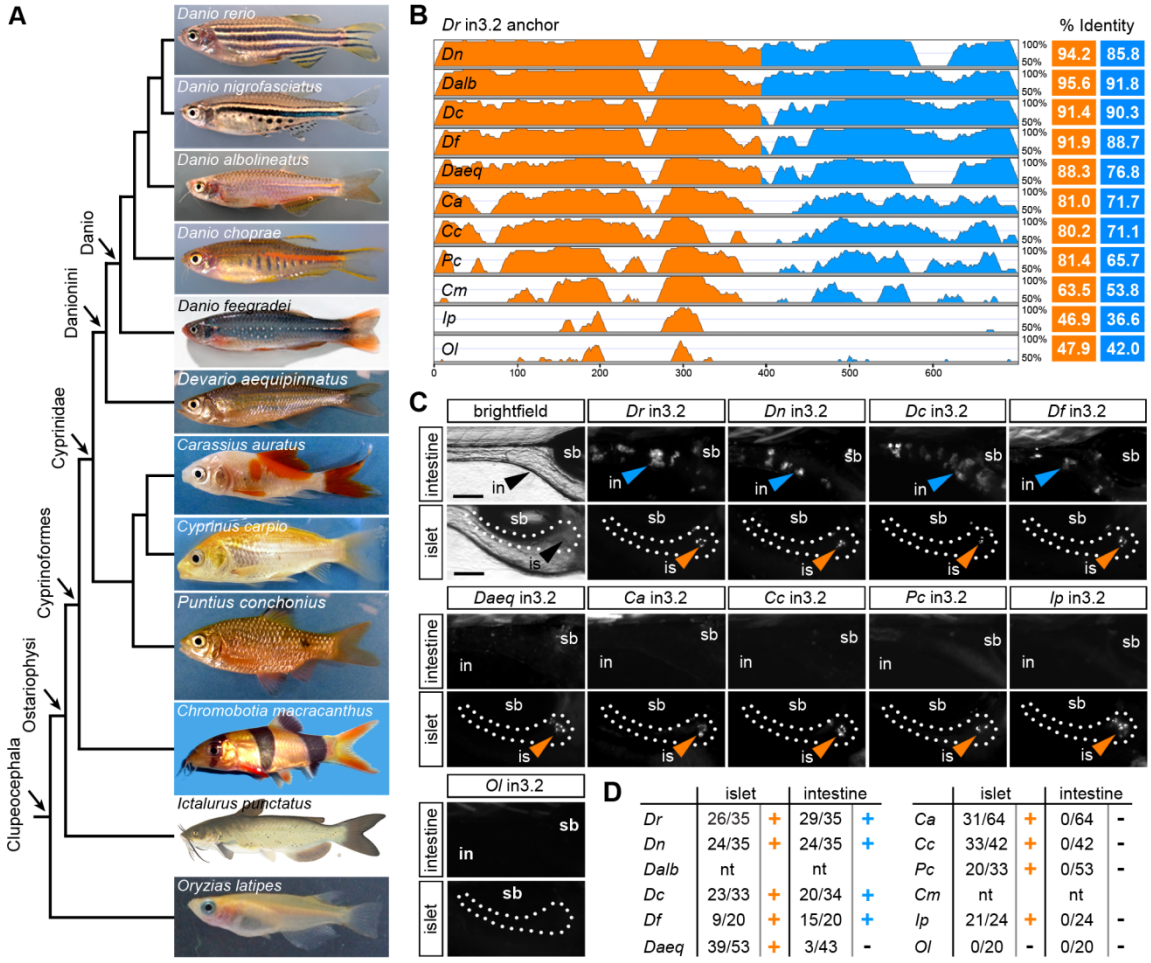


Figure 3.4: Functional evolution of the islet and intestinal regulatory modules in 12 fish species. (A) Unscaled phylogram based on information from [204,205] showing images and relative relationships of 12 fish for which intronic sequences were analyzed. *Danio rerio* (*Dr*, zebrafish), *Danio nigrofasciatus* (*Dn*), *Danio albolineatus* (*Dalb*), *Danio choprae* (*Dc*), *Danio feegradei* (*Df*), *Devario aequipinnatus* (*Daeq*, giant danio), *Carassius auratus* (*Ca*, goldfish), *Cyprinus carpio* (*Cc*, carp), *Puntius conchonius* (*Pc*, rosy barb), *Chromobotia macracanthus* (*Cm*, clown loach), *Ictalurus punctatus* (*Ip*, channel catfish), *Oryzias latipes* (*Ol*, medaka). (B) VISTA plot displaying the global pairwise alignment of orthologous in3.2 regions from each species anchored to zebrafish (*Dr*) in3.2. Orange peaks correspond to regions in the alignment that correspond to *Dr* in3.3 (islet module). Blue peaks correspond to regions in the alignment that correspond to *Dr* in3.4 (intestine module). Percent identity is calculated from pairwise alignments of each module with zebrafish (VISTA parameters: 25 bp sliding window, LAGAN alignment). (C) Representative islet and intestinal images from injections of each orthologous in3.2 module. Orange or blue arrowheads mark positive islet or intestine expression, respectively. The absence of arrowheads denotes negative expression in each tissue. (D) Summary of mosaic expression for each species. Ratios of islet or intestine positive fish versus total fish expressing gfp are shown. Orange or blue (+) denotes that the construct was sufficient to confer expression in the islet or intestine, respectively. Black (-) denotes insufficiency. Note that *Dalb* and *Cm* sequences were not tested (nt) in this heterologous functional assay. See also Figures 3.S5 and 3.S6.

3.3.6 Truncation mapping of the islet and intestinal *angptl4* intronic regulatory modules

Guided by our conservation analyses, we next sought to map the boundaries of critical regulatory regions in the zebrafish in3.3 islet and in3.4 intestinal CRMs by creating and testing truncations of these modules. Each truncation construct was injected into embryos and analyzed at 6-7 dpf for mosaic expression in the islet or intestine. These analyses defined a 164 bp region sufficient to confer islet expression (in3.17; Figure 3.5A,B) including a 129 bp region present in all islet-sufficient truncations (Figure 3.5A). This 129 bp region overlaps with conserved regions identified in our comparative evolutionary analysis (Figure 3.7A). *In silico* prediction of transcription factor binding sites in this critical region identified putative binding sites for multiple transcription factors known to be active in pancreatic islets such as Myc [210,211] and Arnt/HIF1b [212,213], as well ubiquitously expressed transcription factors with important regulatory roles in β-cells such as USF [214] and CREB/ATF [215] (Figure 3.7A).

A distinct 116 bp region (in3.12) was found to be sufficient to confer intestinal expression (Figure 3.5A,C). Notably, the intensity driven by in3.12 in the intestine was lower than other larger truncations of this module that confer strong intestine-specific expression, such as in3.9 and in3.11 (Figure 3.5C,D). The in3.12 truncation therefore represents a minimal intestinal regulatory module that requires additional flanking sequence information to facilitate maximal activity. Intriguingly, the in3.11 truncation, which displays strong intestinal activity, overlaps with two regions of high conservation identified in our comparative evolutionary analysis (Figure 3.7B), suggesting that specific sequences within these conserved regions may be responsible for mediating intestine-specific enhancer activity. Together, these results define the approximate boundaries of functional regulatory DNA within *angptl4* intron 3 required for intestinal and islet transcription.

3.3.7 Site-directed mutagenesis confirms functional motifs within the intestinal module

To complement our comparative genomic and truncation strategies, we used site-directed mutagenesis (SDM) to generate a higher-resolution understanding of the functional DNA motifs required for enterocyte-specific transcription of *angptl4*. Ten base-pair substitutions were tiled across the region corresponding to in3.11 within the context of the entire in3.4 module, and assayed for competency to drive intestinal transcription (Figure 3.6A). This analysis revealed two regions of 40 bp and 20 bp that disrupt intestinal reporter expression when mutated (Figure 3.6B,C). DNA adjacent to these regions was not required for intestinal expression, validating the efficacy of the experimental approach. These data support our truncation mapping experiments (Figure 3.5) by localizing a required region within the in3.12 truncation, as well as a second region within the larger, more active in3.11 truncation. We observed strong overlap between conserved sequences in intestine-positive in3.4 modules identified in our comparative genomic analysis and regions identified by SDM as required for intestinal expression (Figure 3.7B). Specifically, SDM revealed that regions deleted in *Daeq* and *Dn* lineages do not harbor functional motifs required for intestinal expression. Most notably, mutation block 4-7 overlap with the single nucleotide polymorphisms between Devario and Danio species (Figure 3.S6). This region harbors predicted binding sites for transcription factors involved in intestinal epithelial cell biology (Figure 3.7B; see Discussion) that represent attractive candidates for controlling enterocyte-specific *angptl4* transcription.

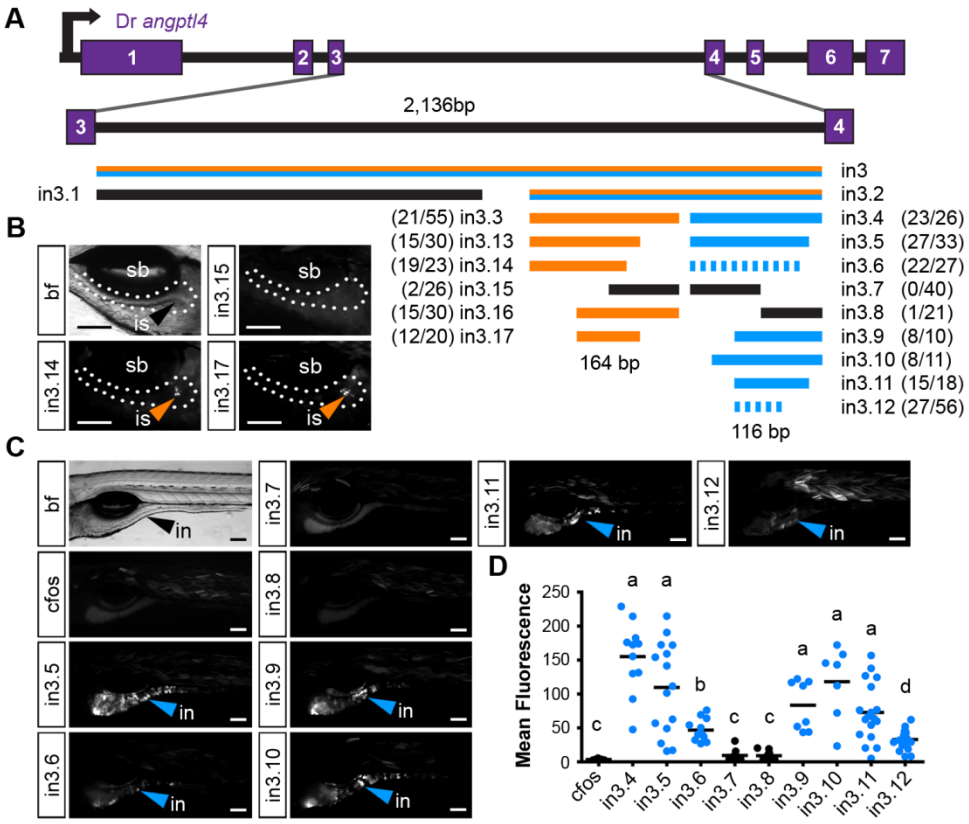


Figure 3.5: Truncation mapping of the islet and intestinal regulatory module.
(A) Scaled schematic of the zebrafish *angptl4* locus showing annotations of truncations assayed for regulatory potential. Orange lines indicate sufficiency to confer islet expression, blue lines indicate sufficiency to confer intestinal expression, and black lines indicate insufficiency in intestine and islet. Dashed blue lines indicate reduced intestinal expression compared to in3.4. Ratios of islet or intestine positive fish versus total fish expressing gfp are shown in parentheses next to truncation labels. (B) Representative images of islet views from mosaic injected fish of each truncation construct. Orange arrows mark islet expression (is). Scale bars = 100 μm. (C) Representative images of intestinal views from mosaic fish injected with each truncation construct. Blue arrows mark intestinal expression (in). Scale bars = 100 μm. (D) Relative mean intestinal fluorescence within the intestine was quantified in mosaic animals (see Materials and Methods) and plotted per injected fish. Circles represent mean fluorescence averaged for three mosaic patches within one fish, and are colored blue or black to designate truncations that are sufficient or insufficient to confer intestinal expression, respectively. Statistical significance was tested using Kruskal-Wallis one-way analysis of variance (labels: a = P<.001, b = P<.05 vs. *Fos*; c = P<.001, d = P<.01 vs. in3.4). Scale bars = 100 μm.

3.3.8 The in3.4 module recapitulates *angptl4* suppression by the microbiota

The presence of commensal gut microbiota in mice results in decreased levels of *Angptl4* transcript specifically in the intestinal epithelium, which is thought to lead to increased peripheral fat storage [8]. However, it remained unknown whether this microbe-induced change in transcript levels was due to alterations in transcriptional activity or transcript stability. We speculated that the intestine-specific *cis*-regulatory module within intron 3 could impart this environmental response in the zebrafish. Our previous comparisons of 6 dpf GF zebrafish to age-matched ex-GF zebrafish colonized since 3 dpf with a normal microbiota (conventionalized or CONVD) indicated that the presence of a microbiota results in reduced *angptl4* transcript levels [90,94,195]. To define the cellular origins of this response in zebrafish hosts, we used semi-quantitative WISH assays to reveal marked reduction of *angptl4* mRNA in the intestinal epithelium in 6 dpf CONVD zebrafish compared to age-matched GF controls (Figure 3.8A). These results indicate that intestinal epithelial suppression of *angptl4* expression is a conserved response to the microbiota in zebrafish and mammalian hosts.

We next tested the ability of the zebrafish intestinal CRM in3.4 to mediate the observed microbial suppression of the endogenous *angptl4* gene. We reared stable *Tg(in3.4-Mmu.Fos:GFP)* zebrafish to 6 dpf under GF or CONVD conditions and assayed transcript levels for both *GFP* reporter and endogenous *angptl4* using qRT-PCR. Consistent with our WISH results, endogenous *angptl4* transcript levels were significantly and reproducibly reduced in CONVD compared to GF animals (Figure 3.8B). Strikingly, transcript levels of the *GFP* reporter gene were similarly reduced in CONVD compared to GF animals (Figure 3.8B). These observations were confirmed using an independent stable transgenic line, *Tg(in3.2-Mmu.Fos:tdT),* harboring the in3.2 reporter which includes the in3.4 module (Figure 3.S7). These data identify the *angptl4*

in3.4 module as a nodal *cis*-regulatory module that integrates transcriptional regulatory input from intestinal epithelial-specific and microbial factors.

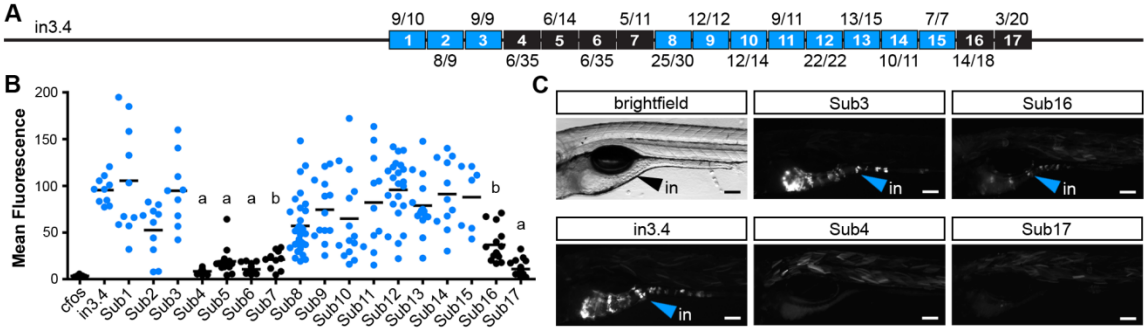


Figure 3.6: Site-directed mutagenesis defines DNA motifs required for intestinal expression.
(A) Scaled schematic showing 10 bp substitution blocks tiled across the zebrafish *angptl4* in3.11 region within the context of the entire in3.4 intestinal module. Black or blue blocks represent mutations that do or do not significantly alter intestinal expression compared to wild type in3.4, respectively (see below). Ratios of intestine positive fish versus total fish expressing GFP are shown in parentheses above or below substitution block labels. (B) Relative mean intestinal fluorescence was quantified in mosaic animals (see Materials and Methods) and plotted per injected fish. Circles represent mean fluorescence averaged for three mosaic patches within a single fish and are colored blue or black to designate mutations that do or do not confer intestinal expression, respectively. Statistical significance was tested using Kruskal-Wallis one-way analysis of variance (labels: a = P<.01 vs. in3.4, P >.05 vs. *Fos*; b = P>.05 vs. *Fos*; unlabeled = P>.05 vs. in3.4, P<.01 vs. *Fos*). (C) Images from animals with mosaic expression of five representative mutant constructs are shown. Blue arrows indicate intestinal expression (in). Scale bars = 100 μm.

3.4 Discussion

3.4.1 Non-overlapping modules confer cell-type specific transcription of *angptl4*

Transcriptional regulation is a key determinant of gene function in the context of animal development and physiology. Recent biochemical and genetic studies in mouse and humans have identified Angptl4 as a critical hormonal regulator of TG-rich lipoprotein metabolism, angiogenesis, and tumor cell survival and metastasis. An improved understanding of the mechanisms controlling Angptl4 activity levels could therefore lead to new approaches for controlling multiple pathophysiologic processes. Although we have a working understanding of Angptl4's post-translational functions, our current knowledge of the mechanisms underlying *Angptl4* transcription in different tissues is relatively limited. Here, we exploited the advantages of the zebrafish model system to examine the regulatory potential of DNA at the *angptl4* locus in all cell types simultaneously and within an intact and living vertebrate organism that can be raised under gnotobiotic conditions. We found that the zebrafish *angptl4* ortholog is expressed in many of the same tissues and cell types as mammalian *Angptl4* (i.e., liver, pancreatic β-cells, and intestinal enterocytes). This finding suggests that the tissue-specific pattern of *Angptl4* expression may have been conserved in the last common ancestor of mammalian and teleost lineages and might have important functional consequences on vertebrate physiology.

Our results reveal that transcription of *angptl4* in distinct tissues might be governed by independent *cis*-regulatory mechanisms. This modular design could have important implications for Angptl4 evolution and function. First, tissue-specific CRMs could have allowed independent evolution of CRM sequence structure. Consistent with this notion, we observed evidence of differential evolution of the islet and intestinal

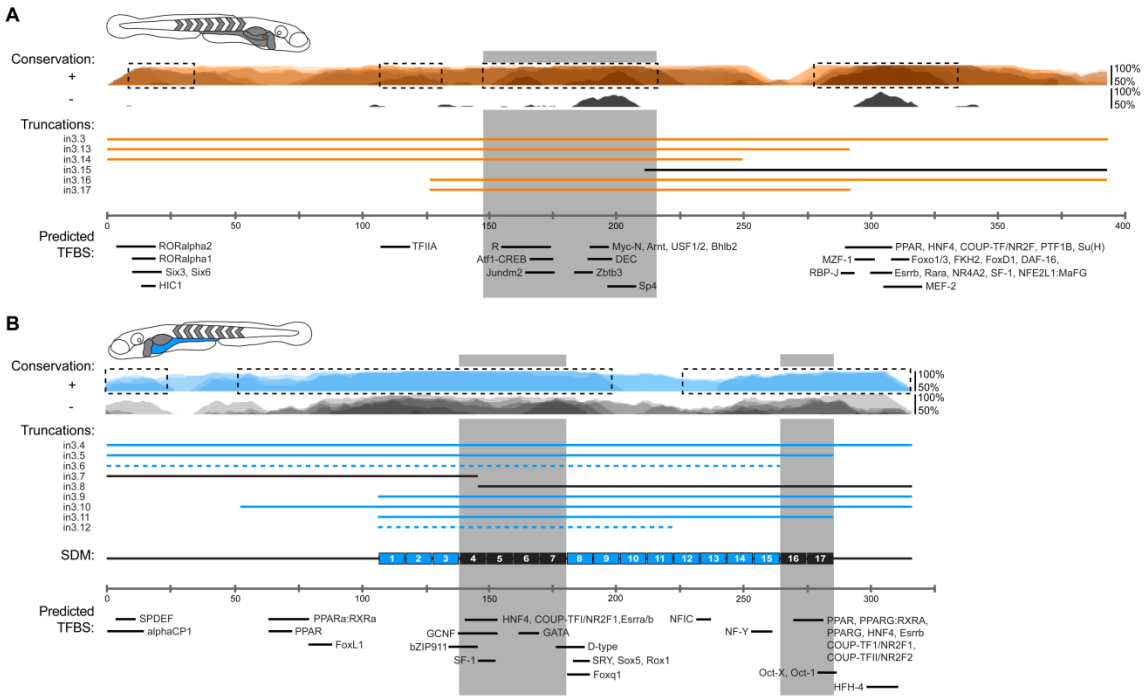


Figure 3.7: Summary of functional conservation and mapping of islet and intestinal regulatory information.
(A) Conservation plots, module truncations, and predicted transcription factor binding sites (TFBS) in islet CRM in3.3 are overlayed and annotated to scale. The grey shaded box represents the region that is present in all positive truncations and has strong conservation in islet-positive species. (B) Conservation plots, module truncations, SDM data, and predicted transcription factor binding sites (TFBS) in intestinal CRM in3.4 are overlayed and annotated to scale. Two grey shaded boxes represent regions that are present in all positive truncations, are required for intestinal expression, and have strong conservation in intestine-positive species. Dotted boxes in panels A and B represent highly conserved regions from each (A) islet-positive or (B) intestine-positive species used to predict common TFBS (see Figures 3.S5 and 3.S6, and Materials and Methods).

modules within teleost fish lineages (Figure 3.4). Differential selective pressures influencing CRM sequence evolution likely arise from the vastly different cellular contexts and exogenous stimuli of each cell type. Pancreatic β-cells are surrounded by other endocrine and exocrine pancreatic cells as well as vascular endothelial cells, whereas intestinal epithelial cells are exposed to complex and variable contents of the intestinal lumen and to the cells of the underlying lamina propria. Combining the observations that (i) functional conservation of the intestinal module is restricted to Danio

species, (ii) transcriptional activity of the intestinal module is sensitive to the microbial status of the intestinal lumen, and (iii) this microbial regulation of *angptl4* transcript levels is conserved in mammals, suggests an intriguing possibility that genes expressed in intestinal epithelia exposed to the dynamic and potentially hazardous luminal environment undergo relatively rapid regulatory evolution. Previous studies have suggested that the expression and function of *defensin* genes within the epithelia of the intestine and other exposed tissues has driven rapid evolution of their coding sequences [216], and our results raise the possibility that similar selective pressures may also affect evolutionary rate of regulatory sequences for *angptl4* and potentially other genes. Second, discrete *cis*-regulatory modules could have led to the independent evolution of Angptl4 synthesis in each respective cell type. This evolution would allow each expressing cell type to independently communicate its physiologic status and environmental exposures systemically by secreting Angptl4 into circulation, and locally by secreting Angptl4 into the extracellular space. The modular organization of these independent tissue-specific CRMs suggests that therapeutic strategies could be developed to control Angptl4 synthesis in specific target tissues without unintended effects on Angptl4 synthesis in other tissues.

Previous studies of CRM evolution in vertebrates and invertebrates have focused primarily on enhancers regulating expression of genes involved in development [196,207,217]. These studies revealed that maintenance of regulatory function can be sustained over long evolutionary distances despite marked sequence dissimilarity and turnover of regulatory information. Our work provides a novel example of utilizing genomic DNA sequences from both close and distant relatives to define the evolutionary dynamics of multiple CRMs and marks the first time to our knowledge that such an extensive exploration (i.e., 12 related fish species) was carried out in a vertebrate. We find that transcriptional output generated by both the intestinal and islet modules is

maintained through a striking conservation in DNA sequence throughout the entire functional module, with little or no turnover of predicted binding sites. This finding suggests that these modules can comply with the "enhanceosome model" of regulatory information organization, as opposed to the "billboard model," which accommodates variation in binding site order, orientation, and spacing [84,218]. However, we detected little non-coding sequence conservation between zebrafish *angptl4* and mouse *Angptl4* intron 3, and we did not detect islet or intestinal reporter expression in a heterologous assay in which we tested full and truncated versions of mouse introns 3 and 4 in the zebrafish (Figure 3.S8 and data not shown). Note that we observed interesting regulatory activity derived from the 3' portion of mouse intron 3 (in3.2; Figure 3.S8C) in enlarged cells circulating within the vasculature in the 1dpf zebrafish. Together these results suggest either that regulatory information governing islet and intestinal expression of murine *Angptl4* is not located within intron 3 or that compensatory *cis*/*trans* mutations render murine intron 3 sequences non-functional in the zebrafish. We suspect that rules governing CRM function and evolution are dependent on the distinct nature of the organism, the specific module, and the signals that the module integrates. It therefore remains an intriguing question as to what extent lessons learned from developmental gene regulation are applicable to the evolution of CRMs controlling expression of genes like *Angptl4* that function in homeostatic physiology or in response to environmental factors like the microbiota [218].

Analyses of Drosophila genomes have elegantly shown that CRM "discovery power scales with the divergence time and number of species compared" [219], and our results suggest that the same will be true in vertebrate lineages. Moreover, our data underscore the need for more reference genome sequences from phylogenetically diverse fish species, in combination with experimentally tractable fish models such as the zebrafish, to facilitate new insights into vertebrate CRM function and evolution.

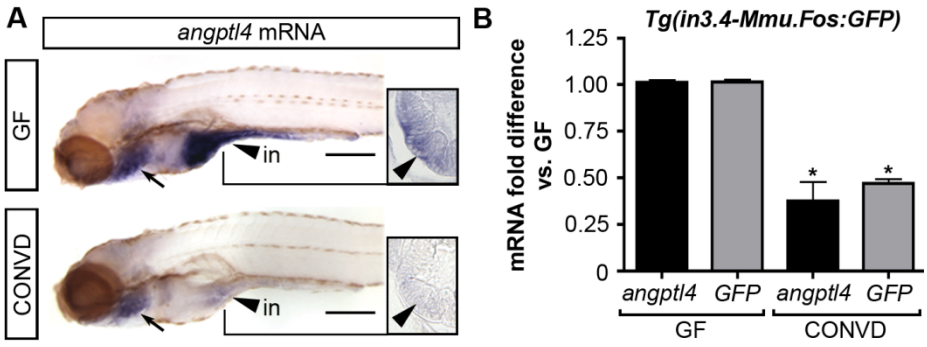


Figure 3.8: The intestinal module in3.4 recapitulates microbial suppression of *angptl4*.
(A) Semi-quantitative whole mount *in situ* hybridization of *angptl4* mRNA in 6 dpf germ-free (GF) and conventionalized (CONVD) animals. Arrowheads mark intestinal expression. Note that the background staining in the gills (arrows) is similar in GF and CONVD fish. Transverse sections show that microbial suppression of *angptl4* mRNA is specific to the intestinal epithelium. (B) Quantitative RT-PCR of *angptl4* and *GFP* mRNA levels in 6 dpf GF and CONVD *Tg(in3.4-Mmu.Fos:GFP)* animals. GF and CONVD animals were derived from the same *Tg(in3.4-Mmu.Fos:GFP)* stable line. *GFP* and *angptl4* mRNA were normalized to *18S* rRNA levels and are shown as fold difference compared to GF controls averaged across 3 experimental replicates ± SEM (2 biological replicate groups of 10 larvae per condition per experiment). Similar results were attained when normalized to *ribosomal protein L32* (*rpl32*) rRNA levels. Asterisks denote P-value <.01 from unpaired T-test between GF and CONVD conditions for each gene. See also Figure 3.S8.

3.4.2 The nature of microbial signals regulating intestinal transcription of *angptl4*

The intestinal microbiota has been identified as an important environmental factor that contributes to host energy storage and obesity, and our results provide critical new insights into how this might be achieved. Previous studies in gnotobiotic mice have shown that the intestinal microbiota regulates fat storage in part by suppressing *Angptl4* transcript levels in the epithelium of the small intestine but not in liver or WAT [8,161]. However, it remained unclear whether these microbe-induced reductions of *Angptl4* transcript levels were due to alterations in *Angptl4* transcription or mRNA turnover. Furthermore, the molecular basis of the intestinal specificity of this response remained unknown. Our results reveal that zebrafish *angptl4* transcript levels are also reduced in the intestinal epithelium in the presence of a microbiota, suggesting that the microbial

regulation of *angptl4* transcript levels might be an evolutionarily ancient feature of host-microbe commensalism in the vertebrate intestine. Our observation that transcript levels from the in3.4 reporter and the endogenous *angptl4* gene respond similarly to microbial colonization strongly suggests that the microbiota regulates *angptl4* expression, at least in part, by reducing the transcriptional activity of this enterocyte-specific enhancer module. These results indicate that enterocyte-specific and microbial control of *angptl4* transcription is conferred through a shared intronic enhancer.

Future investigation will be required to determine whether microbial regulation of in3.4 activity is achieved by (i) reducing the accessibility of this chromatin region to activating *trans*-factors, (ii) subverting the expression or activity of activating *trans*-factors, and/or (iii) inducing expression or activity of repressive *trans*-factors that function through this module. To distinguish between these models, it will be useful to identify the microbial activity and host transcription factors that regulate *angptl4* transcription in the intestinal epithelium. We previously reported that colonization of GF zebrafish with a microbiota harvested from conventionally raised zebrafish or mice resulted in similar suppression of *angptl4* transcript levels in the digestive tract [195]. This finding suggests that the microbial factor(s) regulating zebrafish *angptl4* transcription is expressed by the 'native' zebrafish microbiota and in the 'non-native' and compositionally distinct mouse gut microbiota. Previous studies have identified individual microbial species sufficient to regulate *angptl4* expression in gnotobiotic zebrafish [90,195] and mouse hosts [11,220] as well as in cultured colon cancer cells [181,221], suggesting that reductionist approaches in these microbial species could be used to define the specific factors they utilize to control expression of *angptl4* homologs and other host genes.

3.4.3 Potential transcription factors regulating intestinal transcription of *angptl4*

In this study, we define two minimal regions within the in3.4 CRM that harbor regulatory activity in the intestine and are also conserved within the Danio lineage (Figure 3.7B). Predicted transcription factor binding sites within these regions intimates potential roles for these factors in regulation of *angptl4* tissue-specific transcription and/or microbial suppression. Because sequence-specific transcription factors typically recognize 6-12 bp motifs [222], it is reasonable to assume that multiple factors cooperate to combinatorially regulate intestinal expression through this CRM. The Hnf4 family of fatty acid-regulated nuclear receptors has evolutionarily conserved roles in lipid metabolism [223,224], and Hnf4a is expressed in the intestinal epithelium of zebrafish [99] and mouse [143]. Similarly, GATA factors 4, 5, and 6 are all expressed in the zebrafish [100,225] and mouse [97,98] intestinal epithelium and have proposed roles in regulating epithelial cell differentiation. Notably, *C. elegans* GATA family member *elt-2* has been implicated in mediating intestinal epithelial cell immune responses [102], suggesting that GATA factors could mediate tissue-specific as well as microbial regulatory inputs at *angptl4*. PPAR family members have been identified as key regulators of mammalian *Angptl4* expression in adipocytes and hepatocytes through PPAR responsive elements located in the 5' portion of human *ANGPTL4* intron 3 [177,180], and zebrafish PPARγ [226] and PPARδ [227] homologs are expressed in the larval intestine. The zebrafish *angptl4* locus contains multiple predicted PPRE sites, including several in both the 5' and 3' portion of intron 3 [228]. Most notably, a predicted PPRE was detected within the substitution blocks 16/17 in the intestinal enhancer in3.4 (Figure 3.7B). However, the PPREs within zebrafish *angptl4* intron 3 that display the highest sequence homology to the defined human *ANGPTL4* intron 3 PPRE mapped outside of minimal regions for either intestinal or islet expression within the 5' liver module (data not shown). The location of these PPREs in the 5' region of zebrafish

angptl4 intron 3, combined with the fact that the PPREs discovered in human *ANGPTL4* are also located in the 5' portion of intron 3, suggests that the predicted PPREs within the 3' islet and intestine CRMs of zebrafish *angptl4* could represent novel elements for which functional equivalents have not been identified in mammals.

Although these predicted factors represent candidates for controlling intestine-specific regulation of *angptl4*, databases of predicted TFBSs are incomplete and commonly produce both false-positive and false-negative predictions. Moreover, critical regions identified by SDM might reflect sequences that alter nucleosome positioning or histone modification patterns rather than binding sites for sequence-specific transcription factors. Therefore, we anticipate that unbiased methods for transcription factor discovery will provide the most rigorous approach to an improved understanding of this *cis*/*trans* system. The structure-function analysis of the zebrafish in3.4 intestinal enhancer module reported here was designed to identify sequences critical for intestinal activity. It will therefore be interesting to determine whether exogenous microbial inputs are interpreted through the same or distinct motifs within this CRM and how the endogenous *trans*-acting factors mediating microbial and intestinal regulatory inputs interact to determine transcriptional output.

3.5 Materials and Methods

3.5.1 Zebrafish husbandry

All experiments using zebrafish were performed in wild-type TL or *Tg(ins:CFP-NTR)*^{s892} [206] strains according to established protocols approved by the Animal Studies Committee at the University of North Carolina at Chapel Hill. New stable transgenic lines generated in this study are listed in Table 3.S3. Conventionally raised zebrafish were reared and maintained as described [226]. Production, colonization,

maintenance, and sterility testing of germ-free zebrafish were performed as described [57,94].

3.5.2 Protein sequence analysis

Protein sequences from top BlastP hits to human (*Homo sapiens,* Hs) ANGPTL4 and zebrafish Angptl4 (*Danio rerio,* Dr) were acquired through NCBI or Ensembl and aligned using MUSCLE with default settings [229]. Amino acids highlighted in black represent identical residues in at least 50% of species, whereas amino acids highlighted in grey represent biochemically similar residues (Boxshade). The cleavage recognition sequence and LPL inhibition domain were annotated using information from previous publications [165,230]. The boundaries of the fibrinogen domain were annotated using *in silico* predictions [231,232]. Gaps in the alignment resulting from poorly annotated sequences were manually curated using primary DNA sequence and *in silico* translated using ExPASy [233]. The workflow for inferring phylogenetic relationships was performed at http://mobyle.pasteur.fr/cgi-bin/portal.py. A distance matrix was computed using Phylip 3.67 (Protdist, JTT matrix, default settings), and trees were built using the neighbor-joining method. Bootstrap analysis was performed from 1000 replicates. PHYLIP software and the maximum likelihood probability model [234] using default settings were used to confirm the phylogeny inferred using distance methods. See Table 3.S1 for a complete list of protein sequences used in this study.

3.5.3 DNA sequence analysis

Genomic DNA sequences encompassing 10 kb upstream, including, and 10 kb downstream of the *Angptl4* locus from *Homo sapiens* (GRCh37:19:8419011:8449257:1), *Mus musculus* (NCBIM37:17:33900702:33928520:-1), *Canis familiaris* (BROADD2:20:55933601:55958821:1)*, Danio rerio (*Zv9:2:23312551:23337293),

Oryzias latipes (MEDAKA1:17:6095931:6120384:1), *Takifugu rubripes* (FUGU4:scaffold_212:367815:391593:1), and *Tetraodon nigroviridis* (TETRAODON8:15:3989265:4012887:1) were acquired through Ensembl. 10 kb was chosen as a cutoff because of proximity to neighboring gene loci. Genomic DNA sequence encompassing the *angptl4* locus from *Danio albolineatus* was generously provided by David Parichy (Department of Biology, University of Washington). For species without available genomic sequence, *angptl4* intron 3 regions were PCR amplified from the relevant genomic DNA using a high-fidelity Taq polymerase (Platinum, Invitrogen) and the primers listed in Table 3.S2. PCR products were cloned into TOPO vector pCR2.1 (Invitrogen) prior to sequencing with M13F primers. An EST corresponding to an *angptl4* homolog in *Ictalurus punctatus* (CK419825) was used to design primers targeting exon 3 and exon 4 for PCR amplification of the full-length intron 3. For Cypriniformes species, ESTs EG548328 (*Rutilus rutilus*), DT085020 (*Pimephales promelas*), GH715226 (*Pimephales promelas*), and AM929131 (*Carassius auratus*) were aligned and used to design primers targeting highly conserved regions in *angptl4* exon 3 and exon 4, which we predicted would function for multiple Cypriniformes species. These primers were used to amplify, clone, and sequence the full-length intron 3 from *Cyprinus carpio* and *Chromobotia macracanthus.* Alignment of *Cc*, *Cm*, and *Dr* revealed 100% conservation at the extreme 5' end of the in3.2 module. We used a forward primer targeting in3.2 in combination with a reverse primer targeting exon 4 for cloning of the remaining Cyprinidae species. The bacterial artificial chromosome golwb118_K01 containing the *angptl4* locus from *Oryzias latipes* was provided by Hiroyo Kaneko (Laboratory of Bioresource, National Institute for Basic Biology, Okazaki, Japan). *Carassius auratus*, *Puntius conchonius*, *Cyprinus carpio*, *Devario aequipinnatus*, and *Chromobotia macracanthus* genomic DNA was extracted from the fins of two individuals acquired from commercial suppliers. Genomic DNA from *Ictalurus punctatus* and Danio

species (*Danio nigrofasciatus*, *Danio choprae, Danio feegradei) from one individual were generously provided by Zhanjiang Liu (Department of Fisheries and Allied Aquacultures, Auburn University) and* David Parichy (Department of Biology, University of Washington), respectively. Novel *angptl4* intron 3 sequences generated in this study were deposited in GenBank with accession numbers JN606312-JN606321. Intronic sequences were aligned in mVISTA using LAGAN [235] and visualized using VISTA conservation plots (100bp windows Figure 3.2 and 25bp windows Figure 3.4) [236].

3.5.4 Motif and transcription factor binding site (TFBS) predictions

DNA sequences were queried for predicted transcription factor binding sites deposited in TRANSFAC [237] and JASPAR [238] databases using MATCH [239] and TESS [240] programs using default settings. We used a discriminative motif MEME [241] search to discover motifs common to islet-positive or intestine-positive intronic regions, using sequences orthologous to in3.4 or sequences orthologous to in3.3, respectively, as negative selectors. To determine if MEME motifs were unique to islet- or intestine-positive regions, we used MAST [242] to query islet-negative (*Ol* in.3) or intestine-negative (*Daeq, Ca, Cc, Pc, Cm, Ip, Ol* in3.4) sequences for islet-positive or intestine-positive MEME motifs, respectively. TOMTOM [243] was used to query MEME hits against TRANSFAC and JASPAR databases.

3.5.5 Whole-mount *in situ* hybridization assays

In situ hybridization was performed in whole zebrafish as described [226], except that heads and tails were removed from euthanized 17 dpf animals prior to fixation. Sense and anti-sense riboprobes targeting zebrafish *angptl4* were generated by digesting plasmid fj89c07 in pBK-CMV (NCBI Accession XM_686956) with NotI (sense) or BamHI (anti-sense), and transcribed *in vitro* using T3 (sense; Epicentre) or T7 RNA

polymerase (anti-sense; Epicentre). Sense riboprobes were used in each experiment as a negative control.

3.5.6 Quantitative reverse transcription PCR assays

Total RNA was extracted from groups of 6 dpf whole zebrafish larvae (10 larvae per group, 2 biological replicate groups per condition per experiment, 2 experimental replicates total) using TRIzol Reagent (Invitrogen) or the Qiagen RNeasy (Qiagen) kit using manufacturer's protocol. qRT-PCR was performed as described [94]. Primers used in qRT-PCR assays are listed in Table 3.S2.

3.5.7 Transcription start site and promoter mapping

ESTs at the zebrafish *angptl4* locus were analyzed using UCSC and Ensembl genome browsers. Total RNA was extracted from adult zebrafish intestines and subjected to 5'RACE using the FirstChoice RLM-RACE kit (Ambion), according to the manufacturer's specifications (see Table 3.S2 for primers). Three clones were sequenced and mapped to the zebrafish *angptl4* locus.

3.5.8 Reporter construct cloning

All PCR reactions used for cloning were performed with high-fidelity DNA polymerase (PfuTurbo, Stratagene; Phusion, Invitrogen; Platinum Taq, Invitrogen) and TOP10 chemically competent *E. coli* (Invitrogen). The bacterial artificial chromosome C177A22 containing the zebrafish *angptl4* locus was used as the template for all zebrafish *angptl4* promoter and intronic PCR amplification and cloning. Mouse BAC (RP24-294G12, CHORI), Medaka BAC (golwb118_K01), and sequenced pCR2.1 clones (*lp, Pc, Cc, Ca, Daeq, Df, Dc, Dn*) containing intronic regions orthologous to zebrafish in3.2 from each species were used as source material for cloning in heterologous

reporter assays. The plasmid pT2cfosGW [186] was used as the vector backbone for all Tol2 transgenic reporter assays. The *Fos* minimal promoter and *angptl4* 5' upstream regions were PCR amplified and directionally cloned into pT2cfosGW using XhoI and BamHI restriction sites. This step removed both the original *Fos* promoter and the upstream Gateway site. Of note, we observed significant levels of reporter expression in muscle tissue upon removal of the Gateway cloning site (Figure 3.5C,D and data not shown). Intronic DNA was cloned upstream of the *Fos* minimal promoter in pT2cfosGW using Gateway reagents as described [186]. The intronic module in3.4 was non-directionally cloned into *Tg(-1kbangptl4:GFP)* using the single BglII site located downstream of SV40polyA. A vector (*Tg(in3.4-Mmu.Fos:GFP)*) containing the *angptl4* intronic module in3.4 was used as the source vector for site-directed mutagenesis. To create site-directed substitutions, 50 bp complementary primers containing two 20 bp regions complementary to in3.4, separated by a 10 bp substitution block, were used in circular PCR followed by DpnI treatment to digest methylated parent plasmid. A ClaI restriction site was incorporated into the 10 bp region in order to screen for mutant bacterial colonies. Selection of nucleotide exchange was generally A-C and G-T, except in cases that would create a site amenable to DamI methylation. All plasmids were verified by Sanger dideoxy terminator sequencing. All primers used are listed in Table 3.S2.

3.5.9 Injections, imaging, and reporter quantification

Co-injections of Tol2 plasmid and transposase mRNA were performed as described [186]. Generally, 100-200 zebrafish embryos were injected at the 1-2 cell stage with approximately 69 pg of plasmid DNA at a DNA:transposase ratio of 1:2. Injections of each construct were performed with at least two sequence-verified plasmids in two independent experiments. Mosaic expression patterns were quantified as follows:

at least 200 fish were visually observed, and at least 10 were scored per construct for positive/negative expression in selected tissues. At least 7-20 fish/construct were imaged at the same magnification and exposure time and densitometric measures were quantified in 8-bit grey scale images using ImageJ software [244]. Three mosaic patches within a given tissue of an imaged fish were quantified for mean fluorescence intensity and averaged. Statistical significance was analyzed using Kruskal-Wallis one-way analysis of variance and Dunn's multiple comparison test using GraphPad Prism software. Injected larvae were raised to adulthood and screened for stable germ-line insertion. Where indicated, patterns identified in mosaic animals were verified in a least two independent stable germ-line insertions (Table 3.S3). In each case, independent pedigrees of the same Tol2 vector displayed the same specific pattern of expression in the intestine, liver, and islet, respectively.

3.5.10 Immunohistochemistry

Staining of fixed and sectioned 6 dpf zebrafish was performed exactly as described [94]. Primary antibodies used in this study were anti-GFP (Rabbit, 1:500, Invitrogen), 2F11 (mouse, 1:200), 4E8 (mouse, 1:200; gifts from Julian Lewis), and secondary antibodies were AF568 (goat anti-mouse, 1:500, Invitrogen) and AF488 (goat anti-mouse, 1:500, Invitrogen).

3.5.11 DNase I hypersensitivity

Three intestines were dissected from adult zebrafish at 1 year post-fertilization, splayed, and washed extensively with 1x PBS. Intestines were incubated for 15 minutes on ice in 5 ml of Dissociation Reagent 1 (1x PBS, 30 mM EDTA, 1.5 mM DTT, 1X Complete protease inhibitors; Roche), then transferred to Dissociation Reagent 2 (1x PBS, 30 mM EDTA, 1x Complete protease inhibitors) and shaken at 25 ^{o}C until epithelial

layers were sufficiently sloughed. Epithelial cells were collected, washed in 1x PBS, and re-suspended in 500 microliters of RSB (10 mM Tris-HCl pH 7.4, 10 mM NaCl, 3mM $MgCl_2$). Cells were gently lysed in 10 ml cold RSB plus 0.075% NP-40 and nuclei pelleted at 500 x G at 4 ^oC for 10 minutes. Nuclei were incubated with various concentrations of Dnase I (0 – 1.5 units, NEB) for 10 minutes at 37 ^oC. Reactions were stopped by adding an equal volume of 2x Lysis Buffer (1% SDS, 200 mM NaCl, 10 mM EDTA, 20 mM Tris pH 7.5, 0.4 mg/ml proteinase K) and incubated overnight at 37 ^oC. Digested DNA was extracted using phenol/cholorform/isoamyl alcohol (Fisher), precipitated with ethanol and sodium acetate, and quantified using a fluorimeter (Qubit, Invitrogen). Quantitative PCR was performed as described above using primers listed in Table 3.S2.

3.6 Acknowledgements

The authors are grateful to Zhanjiang Liu for providing *Ictalurus punctatus* gDNA*; to* David Parichy for providing *Danio* spp. gDNA and images; to Shinichi Morishita for the medaka image; to New York State Department of Environmental Conservation for the channel catfish drawing; to Tom Randall for assistance with inferring phylogenies; to Kirk McNaughton for assistance with cryosectioning; to Michele Kanther, James Minchin, Neal Kramarcy for assistance with confocal microscopy; to Christopher Bryant for biostatistics support; and to Scott Bultman and Scott Magness for helpful comments on the manuscript.

3.7 Supporting Information

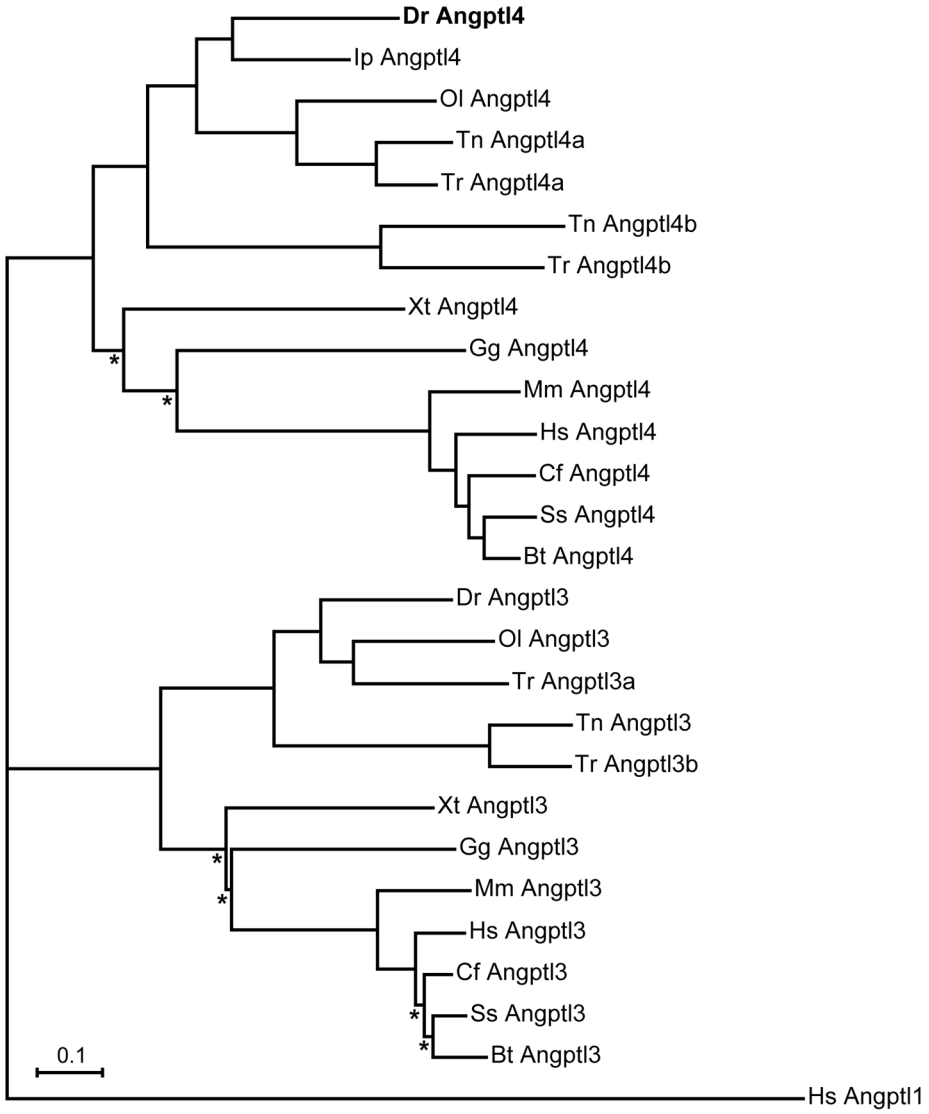


Figure 3.S1: Phylogeny of Angptl4 and Angptl3 proteins from multiple vertebrate species. Distance phylogram of Angiopoietin-like 3 and 4 from zebrafish (*Dr*, *Danio rerio*), catfish (*Ip*, *Ictalurus punctatus*), medaka (*Ol*, *Oryzias latipes*), tetraodon (*Tn*, *Tetraodoan nigroviridis*), fugu (*Tr*, *Takifugu rubipres*), xenopus (*Xt*, *Xenopus tropicalis*), chicken (*Gg*, *Gallus gallus*), mouse (*Mm*, *Mus musculus*), human (*Hs*, *Homo sapiens*), dog (*Cf*, *Canis familiaris*), pig (*Ss*, *Sus scrofa*), and cow (*Bt*, *Bos taurus*). All nodes are significant (>700/1000 bootstrap replicates) except those marked with an asterisk (*). Phylogenic relationships inferred through Maximum Likelihood yield similar branching with differences only in the positions of the nodes separating Xt Angptl3 and Angptl4 and Gg Angptl3 and Angptl4 from mammals (data not shown). Scale bar indicates phylogenetic distance, in number of amino acid substitutions per site. See Table 3.S1 for protein sequences.

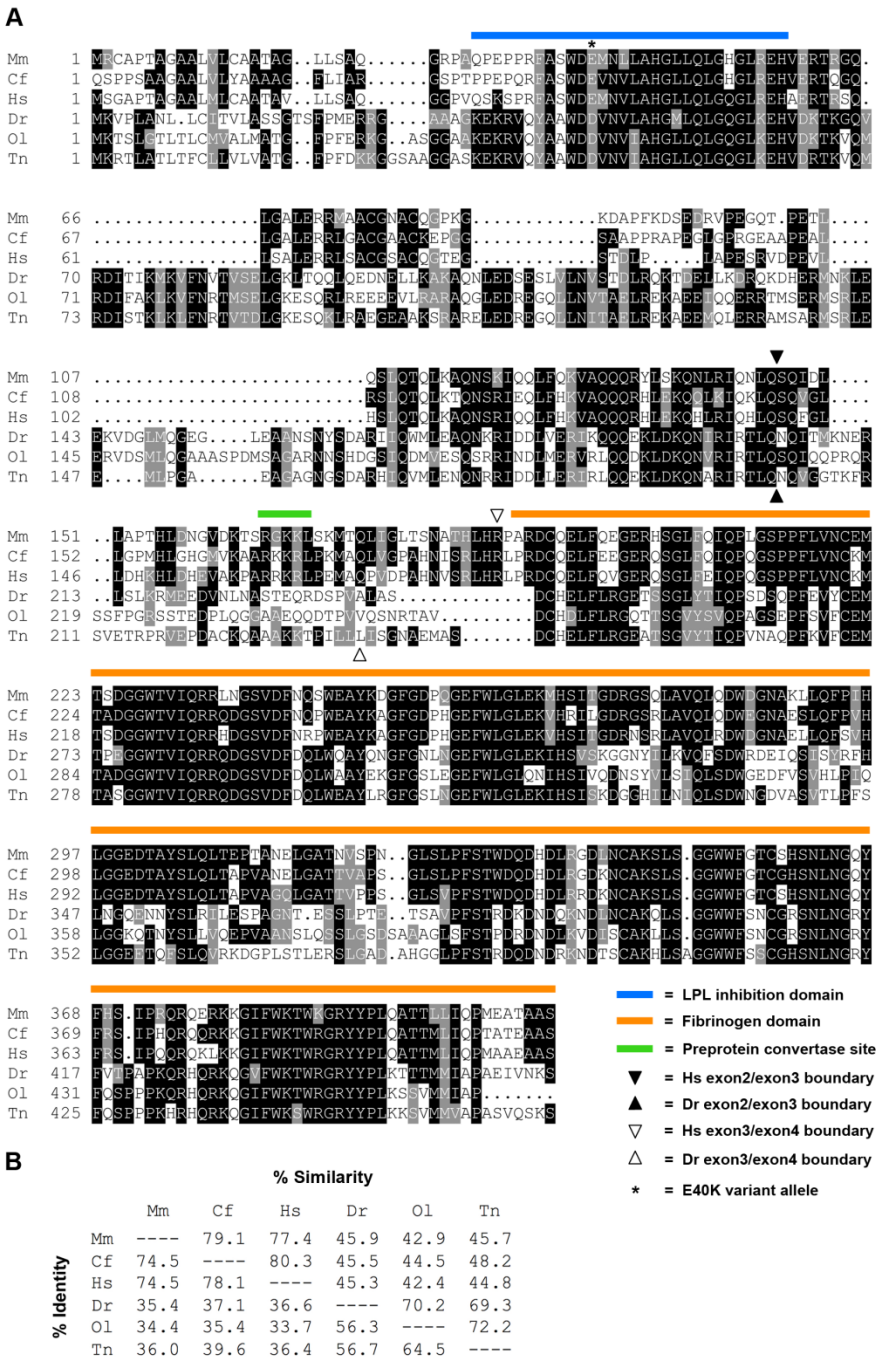


Figure 3.S2: Alignment of Angptl4 proteins from multiple vertebrate species.
(A) Multiple sequence alignment of Angptl4 proteins from representative vertebrate species. Amino acids highlighted in black represent identical residues in at least 50% of species, whereas amino acids highlighted in grey represent biochemically similar residues. The green line denotes the cleavage recognition sequence [230], the blue line denotes the experimentally defined LPL inhibition domain [164], and the orange line denotes the *in silico* predicted fibrinogen domain. Black downward arrows designate the exon 2/3 boundary in human, black upward arrows designate the exon2/3 boundary in zebrafish. White downward arrows designate the exon 3/exon 4 boundary in human, white upward arrows designate the exon 3/exon 4 boundary in zebrafish. The black asterisk marks the position of the human E40K variant [166]. (B) Percent identity and percent similarity matrix for each species pair.

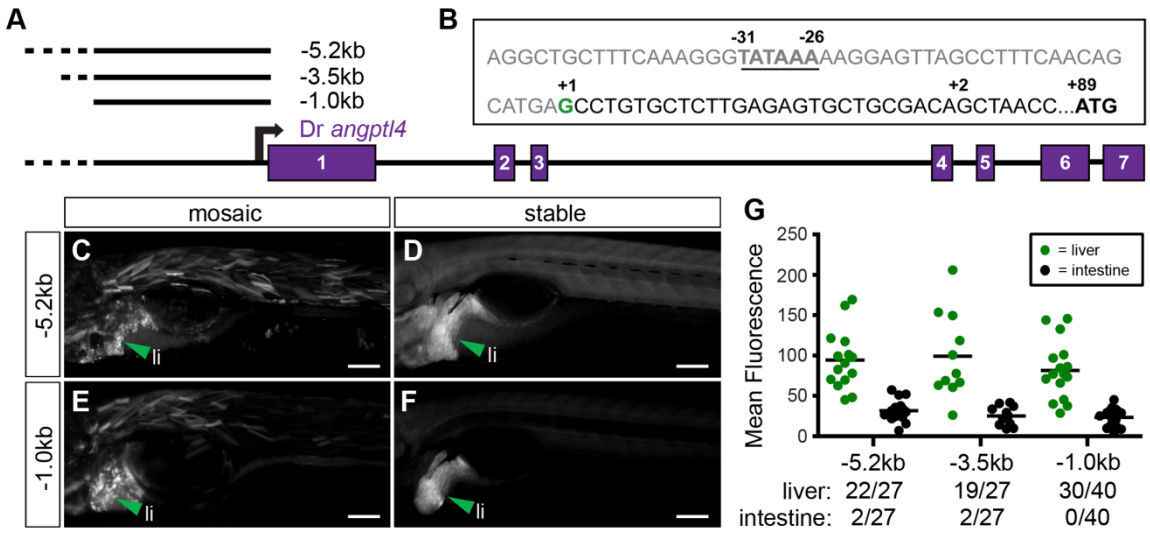


Figure 3.S3: Non-coding DNA upstream of the zebrafish *angptl4* transcription start site drives expression in the liver but not in the intestine or islet.
(A) The zebrafish *angptl4* locus and positions of promoter regions assayed in 0-7 dpf transgenic zebrafish are annotated to scale. (B) 5' RACE and EST data (not shown) establish a single transcription start site directly upstream of exon 1. The positions of the TATA box, transcription start site, and translation start site are annotated. (C, E) Non-coding DNA -5.2 kb and -1 kb upstream of the translation start site drives expression in the liver in 6 dpf mosaic animals. Note that the -5.2 kb fragment includes a region -4.9 kb upstream from the TSS that shares extensive homology with medaka (see Figure 3.2A). Scale bars = 50 μm. (D, F) Liver expression pattern is confirmed in the F_1 generation of injected animals harboring stable insertions of the -5.2 kb (*Tg(-5.2angptl4:GFP)*) and -1kb transgenes (*Tg(-1angptl4:GFP)*). Scale bars = 50 μm. (G) Fluorescence intensity in mosaic animals is quantified (see Materials and Methods) in the liver and intestine. Circles represent mean fluorescence averaged in three mosaic patches within the liver (green) or intestine (black) of 1 fish. Note that there is minimal to no reporter expression in either the intestine or the islet (not shown). Ratios of liver or intestine positive fish versus total fish expressing GFP are shown below the corresponding construct name.

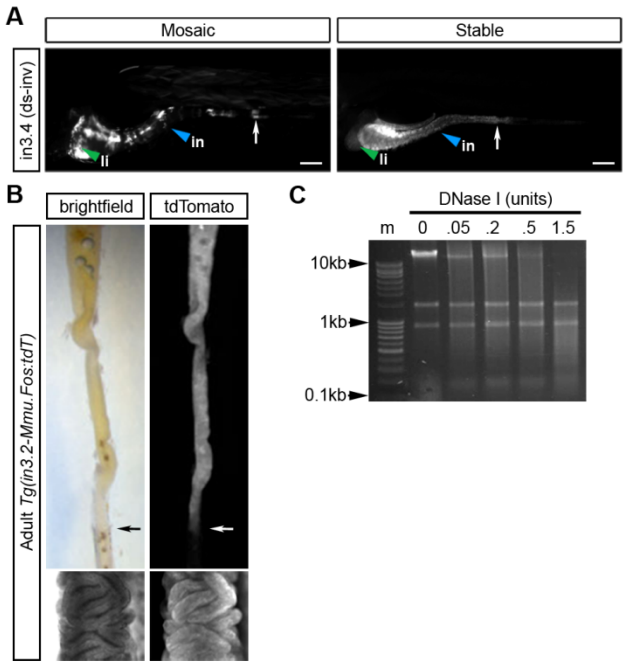


Figure 3.S4: The zebrafish *angptl4* in3.4 intestinal module exhibits hallmarks of a classical enhancer. (A) *Dr* in3.4 was cloned in an inverted orientation (in3.4(ds-iv)) downstream of *GFP* driven by -1 kb of the *angptl4* promoter (*Tg(-1angptl4:GFP:in3.4inv)*). Mosaic and stable intestinal expression patterns are indistinguishable from those when in3.4 is upstream of the *Fos* minimal promoter (see Figure 3.3). The white arrow marks the boundary between the anterior intestine (segment 1) and mid-intestine (segment 2). The marked liver expression is likely conferred by the -1 kb *angptl4* promoter (see Figure 3.S3F). (B) The in3.2 module drives expression of a reporter (tdTomato) in the intestinal epithelium of adult zebrafish. (C) Nuclei were isolated from adult zebrafish epithelial cells and subjected to increasing concentrations of DNase I. Digested DNA from 0.5 units DNase I was used for quantitative PCR shown in Figure 3.3P.

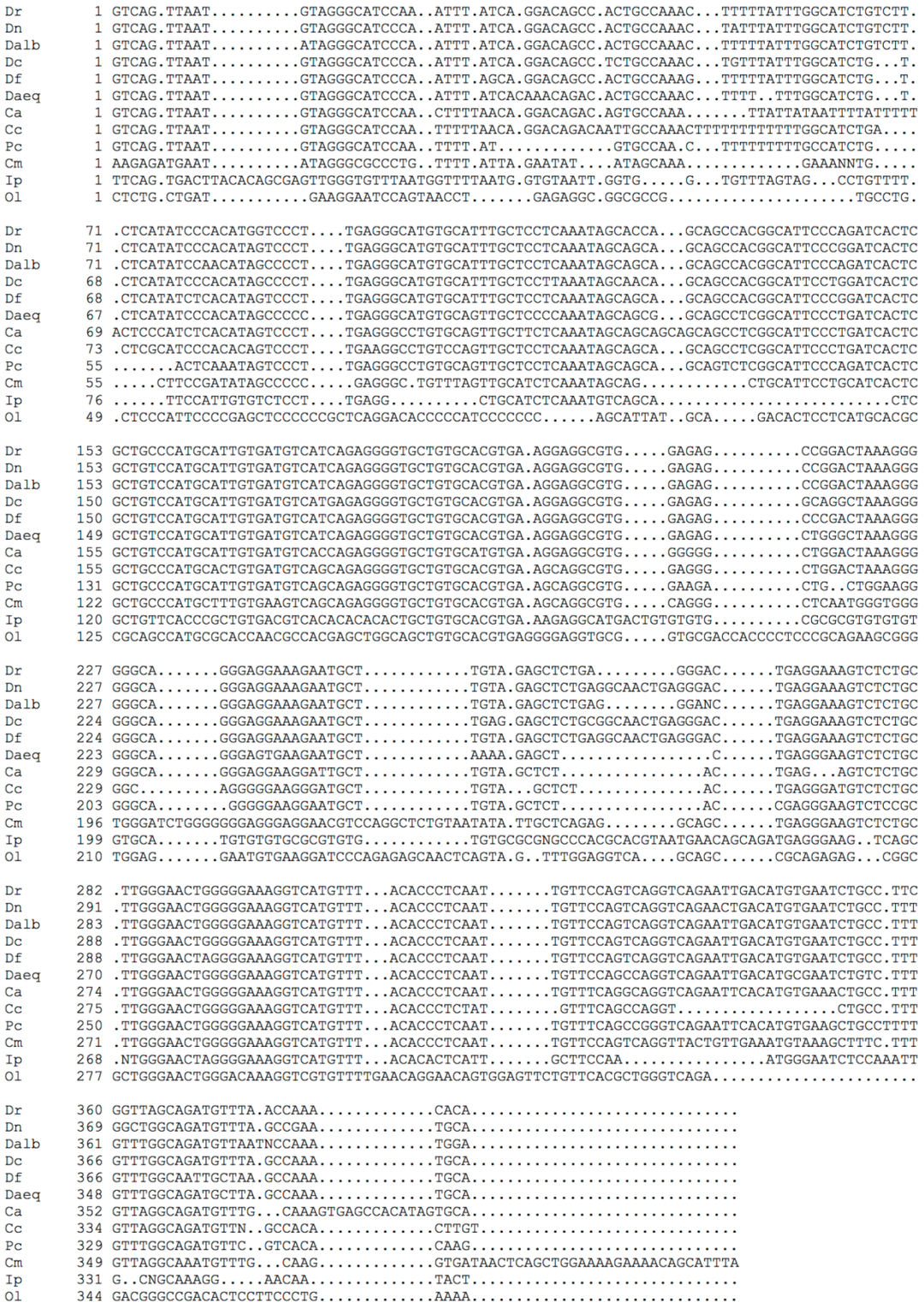


Figure 3.S5: Multiple-species sequence alignment of teleost *angptl4* in3.3 modules.
Sequence alignment (MUSCLE) of in3.3 regions from 12 teleost species.

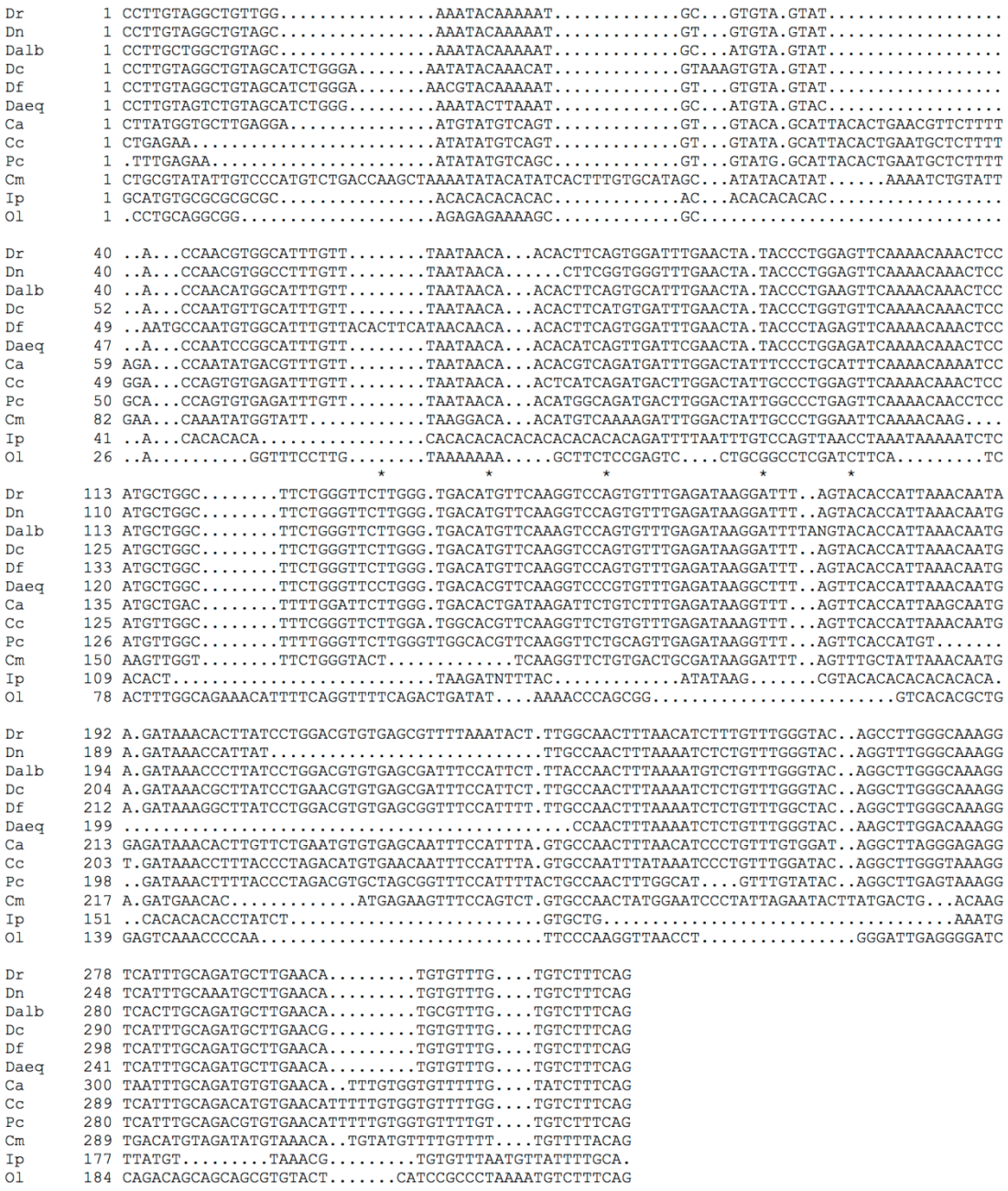


Figure 3.S6: Multiple-species sequence alignment of teleost *angptl4* in3.4 modules.
Sequence alignment (MUSCLE) of in3.4 regions from 12 teleost species. Asterisks mark 5 individual bp changes that are differentially conserved in intestine-positive modules versus intestine-negative modules within the critical region defined by truncation mapping and SDM.

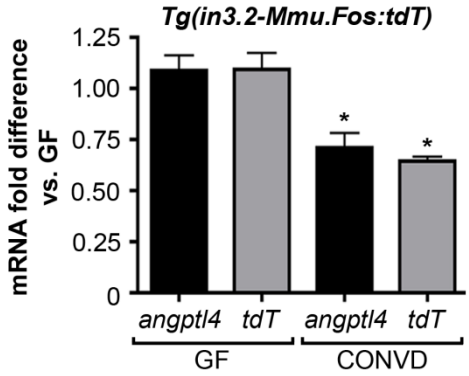


Figure 3.S7: The intronic module in3.2 recapitulates microbial suppression of *angptl4*.
Quantitative RT-PCR of *angptl4* and *tdT* in dissected digestive tracts from 6 dpf GF and CONVD *Tg(in3.2-Mmu.Fos:tdT)* animals. GF and CONVD animals were derived from the same *Tg(in3.2-Mmu.Fos:tdT)* stable line. *tdT* and *angptl4* mRNA were normalized to *18S* rRNA levels and are shown as fold difference compared to GF controls averaged across 3 experimental replicates ± SEM (3 biological replicate groups of 10 digestive tracts per condition per experiment). Asterisks denote P-value <.05 from unpaired T-test between GF and CONVD conditions for each gene. Note that module in3.2 includes the intestinal module in3.4 (see Figure 3.3).

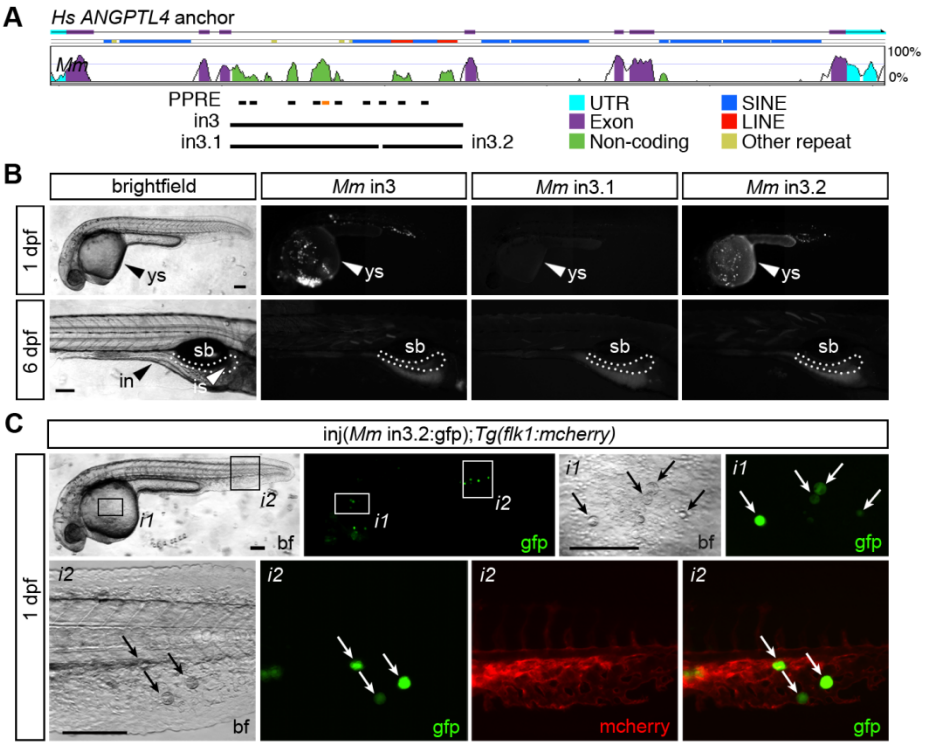


Figure 3.S8: Mouse intron 3 drives expression in circulating blood cells but not in the zebrafish liver, islet, or intestine. (A) VISTA plot displaying the global pairwise alignment of the human *ANGPTL4* locus with the orthologous mouse region. Purple conservation peaks correspond to exons and green conservation peaks represent putative non-coding regions. The 5' and 3' untranslated regions (UTR) are teal. Repeat regions are annotated above the plot. Mouse regions assayed for regulatory potential are indicated as lines below the plot. Predicted PPAR response elements (PPRE) are annotated with the orange line representing an experimentally verified PPRE. (B) Representative images from 1dpf and 6dpf zebrafish injected with each mouse intronic region. There is strong reporter expression in circulating blood cells in 1 dpf fish driven by full-length in3 and the in3.2 truncation. There is no expression in the liver, islet or intestine in 6 dpf animals from any of the constructs assayed. (C) The in3.2 reporter construct was injected into zebrafish harboring the Flk1:mcherry transgene. Note that GFP expression is robust in circulating blood cells that appear enlarged compared to non-GFP expressing cells. Enlarged circulating blood cells have not been observed in any other construct tested.

Table 3.S1: Angiopoietin-like protein sequences used for inferring phylogeny.

>Hs_Angiopoietin-like 1_NP_004664
MKTFTWTLGVLFFLLVDTGHCRGGQFKIKKINQRRYPRATDGKEEAKKCAYTFLVPEQRITGPICVNTKGQD
ASTIKDMITRMDLENLKDVLSRQKREIDVLQLVVDVDGNIVNEVKLLRKESRNMNSRVTQLYMQLLHEIIRKR
DNSLELSQLENKILNVTTEMLKMATRYRELEVKYASLTDLVNNQSVMITLLEEQCLRIFSRQDTHVSPPLVQV
VPQHIPNSQQYTPGLLGGNEIQRDPGYPRDLMPPPDLATSPTKSPFKIPPVTFINEGPFKDCQQAKEAGHSV
SGIYMIKPENSNGPMQLWCENSLDPGGWTVIQKRTDGSVNFFRNWENYKKGFGNIDGEYWLGLENIYMLS
NQDNYKLLIELEDWSDKKVYAEYSSFRLEPESEFYRLRLGTYQGNAGDSMMWHNGKQFTTLDRDKDMYA
GNCAHFHKGGWWYNACAHSNLNGVWYRGGHYRSKHQDGIFWAEYRGGSYSLRAVQMMIKPID
>Bt_ Angiopoietin-like 3_NP_001073814.1
MYTIKLFLIIAPLVISSRTDQDYTSLDSISPEPKSRFAMLDDVKILANGLLQLGHGLKDFVHKTKGQINDIFQKLN
IFDQSFYDLSLQTNEIKEEEKELRRATSKLQVKNEEVKNMSLELDSKLESLLEEKILLQQKVRYLEDQLTDLIK
NQPQIQEYLEVTSLKTLVEQQDNSIKDLLQIVEEQYRQLNQQQSQIKEIENQLRRTGIKESTEISLSSKPRAPR
TTPSFHSNETKNVEHDDIPADCTIIYNQGKHTSGIYSIRPSNSQVFNVYCDVKSGSSWTLIQHRIDGSQNFNE
TWENYKYGFGRLDGEFWLGLEKIYSIVMQSNYILRIELEDWKDKYYTEYSFHLGDHETNYTLHLAEISGNGP
KAFPEHKDLMFSTWDHKAKGHFNCPESNSGGWWYHDVCGENNLNGKYNKPKAKAKPERKEGICWKSQD
GRLYSIKATKMLIHPSDSENSE
>Bt_ Angiopoietin-like 4_NP_001039508.1
MRCAPTAGAALMLCAATAGLLSAQGRPEPPETPRFASWDEVNVLAHGLLQLGHGLREHVERTRGQLGELE
RRLGACGAACKDPEGSAAPPRAQANLVNPGGGDASPETLRSLKTQLEAQNSRIQQLFQKVAQQQRHLEKQ
QLRIQNLQSQMDHLAPRHLGHEMAKPARRKRLPKMAQLAGPAHNISRLHRLPRDCQELFEEGERESGLFQI
QPQGSPPFLVNCKMTSDGGWTVIQRRQDGSVDFNQPWEAYKDGFGDPQGEFWLGLEKVHHILGDRGSRL
AVQLQDWEGNAESLQFPIHLGGEDTAYSLQLTPPVASKLGATTFSPSGLSLPFSTWDQDHDLRGDKNCAR
SLSGGWWFGTCSHSNLNGQYFHSIPRQRQQRKKGIFWKTWRGRYYPLQATTILVQPTAAS
>Cf_ Angiopoietin-like 3_ENSCAFP00000027734
MYTIKLFLFIIPLVISSKIDRDYSSYDSVSPEPKSRFAMLDDVKILANGLLQLGHGLKDFVHKTKGQINDIFQKLN
IFDQSFYDLSLQTNEIKEEEKELRRTTSKLQVKNEEVKNMSLELNSKVESLLEEKILLQQKVRYLEKQLTSLIK
NQPEIQEHPEVTSLKTFVEQQDNSIKDLLQTVEEQYRQLNQQHSQIKEIENQLRNVIQESTENSLSSKPRAPR
TTPFLHLNETKNVEHNDIPANCTTIYNRGEHTSGIYSIRPSNSQVFNVYCDVKSGSSWTLIQHRIDGSQNFNE
TWENYRYGFGRLDGEFWLGLEKIYSIVKQSNYILRIELEDWNDNKHYIEYFFHLGNHETNYTLHLVEITGNILN
ALPEHKDLVFSTWDHKAKGHVNCPESYSGGWWWHNVCGENNLNGKYNKQRAKTKPERRRGLYWKSQN
GRLYSIKSTKMLIHPIDSESSE
>Cf_ Angiopoietin-like 4_XP_533928.3
QSPPSAAGAALVLYAAAAGFLIARGSPTPPEPQRFASWDEVNVLAHGLLQLGHGLREHVERTQGQLGALER
RLGACGAACKEPGGSAAPPRAPEGLGPRGEAAPEALRSLQTQLKTQNSRIEQLFHKVAQQQRHLEKQQLKI
QKLQSQVGLLGPMHLGHGMVKAARKKRLPKMAQLVGPAHNISRLHRLPRDCQELFEEGERQSGLFQIQPQ
GSPPFLVNCKMTADGGWTVIQRRQDGSVDFNQPWEAYKAGFGDPHGEFWLGLEKVHRILGDRGSRLAVQ
LQDWEGNAESLQFPVHLGGEDTAYSLQLTAPVANELGATTVAPSGLSLPFSTWDQDHDLRGDKNCAKSLS
GGWWFGTCGHSNLNGQYFRSIPHQRQQRKKGIFWKTWRGRYYPLQATTMLIQPTATEAAS
>Dr_ Angiopoietin-like 3_NP_571893.1
MLILLLWLSLSTTSAAPNSKKSPTEAPILITAPPTEARSRFAMLDDVRLLANGLLQLGQSLREFVHKTKSQINGI
FQKLNVFDRSFYQLSVVTSEIKEEEEKLKETTIFLKANNEEIRNLSLEINSKINNILQERSQLHTKVGGLEEKLK
GLSQSMMPLEQLQEITALKDVIETQERTITDLLRSVKEQHDQLNYQKIKIKSLEDKVNYDTFQDTIEKPMDLNP
ETPDPFLYLTTNSTNGTKDINDFPADCSEVFTRGQKTSGIYPIKPNQSEPFYVYCEITPDGAATVIQRREDGS
VDFDQSWEKYEHGFGKLEKEFWLGLAKIHSIAQQGEYILHIELEDWKEEKRFIEYTFTLEGPASDYALHLAPL
SGDLSDAMSNHTGMKFSTKDRDNDNHDESNCARNYTGGWWFDACGDTNLNGRYAWMRSKARHQRRKG
SSYTLKSTKITIRPSTHFNNP
>Dr_ Angiopoietin-like 4_NP_001243132.1
MKVPLANLLCITVLASSGTSFPMERRGAAAGKEKRVQYAAWDDVNVLAHGMLQLGQGLKEHVDKTKGQVR
DITIKMKVFNVTVSELGKLTQQLQEDNELLKAKAQNLEDSESLVLNVSTDLRQKTDELLKDRQKDHERMNKL
EEKVDGLMQGEGLEAANSNYSDARIIQWMLEAQNKRIDDLVERIKQQQEKLDKQNIRIRTLQNQITMKNERL
SLKRMEEDVNLNASTEQRDSPVALASDCHELFLRGETSSGLYTIQPSDSQPFEVYCEMTPEGGWTVIQRRQ
DGSVDFDQLWQAYQNGFGNLNGEFWLGLEKIHSVSKGGNYILKVQFSDWRDEIQSISYRFHLNGQENNYS

LRILESPAGNTESSLPTETSAVPFSTRDKDNDQKNDLNCAKQLSGGWWFSNCGRSNLNGRYFVTPAPKQR
HQRKQGVFWKTWRGRYYPLKTTTMMIAPAEIVNKS
>Hs_ Angiopoietin-like 3_NP_055310.1
MFTIKLLLFIVPLVISSRIDQDNSSFDSLSPEPKSRFAMLDDVKILANGLLQLGHGLKDFVHKTKGQINDIFQKL
NIFDQSFYDLSLQTSEIKEEEKELRRTTYKLQVKNEEVKNMSLELNSKLESLLEEKILLQQKVKYLEEQLTNLI
QNQPETPEHPEVTSLKTFVEKQDNSIKDLLQTVEDQYKQLNQQHSQIKEIENQLRRTSIQEPTEISLSSKPRA
PRTTPFLQLNEIRNVKHDGIPAECTTIYNRGEHTSGMYAIRPSNSQVFHVYCDVISGSPWTLIQHRIDGSQNF
NETWENYKYGFGRLDGEFWLGLEKIYSIVKQSNYVLRIELEDWKDNKHYIEYSFYLGNHETNYTLHLVAITGN
VPNAIPENKDLVFSTWDHKAKGHFNCPEGYSGGWWWHDECGENNLNGKYNKPRAKSKPERRRGLSWKS
QNGRLYSIKSTKMLIHPTDSESFE
>Hs_ Angiopoietin-like 4_NP_647475.1
MSGAPTAGAALMLCAATAVLLSAQGGPVQSKSPRFASWDEMNVLAHGLLQLGQGLREHAERTRSQLSALE
RRLSACGSACQGTEGSTDLPLAPESRVDPEVLHSLQTQLKAQNSRIQQLFHKVAQQQRHLEKQHLRIQHLQ
SQFGLLDHKHLDHEVAKPARRKRLPEMAQPVDPAHNVSRLHRLPRDCQELFQVGERQSGLFEIQPQGSPP
FLVNCKMTSDGGWTVIQRRHDGSVDFNRPWEAYKAGFGDPHGEFWLGLEKVHSITGDRNSRLAVQLRDW
DGNAELLQFSVHLGGEDTAYSLQLTAPVAGQLGATTVPPSGLSVPFSTWDQDHDLRRDKNCAKSLSGGW
WFGTCSHSNLNGQYFRSIPQQRQKLKKGIFWKTWRGRYYPLQATTMLIQPMAAEAAS
>Gg_ Angiopoietin-like 3_NP_001128594.1
MKIILLLLFVAPLALSVRAEKDFAFLDSAATPETKSRFAMLDDVRILANGLLQLGHGLKDFVHKTKGQMNDIFQ
KLYIFDRSFYELSLQTSEIKEEEEQLRQTTARLQINNEEIKNLSQEMNLKIEDLIQNKIQLQEKVWGLEDKVTKL
AIIQPTVQETNEISSLKAFVEQQDNHIKQLHKVVEDQHVQLDKQHNQIMELEDKLNHIELQELAENSFLEEQA
ESNEGSPFLVHNSTAVMHKLEGATPDCTALYNSGIRSSGIYTIKPNGSEAFDVYCEMKFGTSWTVIQNRVDG
SLDFNQTWDAYTNGFGDLNEEFWLGLNKTFSITKQGDYILRIELQDWKDNKRYVEYAFTLGGPETDYVLQLS
RISGSIPNALPEQTELRFSTADRDMAIINDLDCPQNYLGGWWHSECEETNLNGKYVTPRSKGRLDRTKGLY
WKPKNGRYYLLKSTKIMIHPTDLKIFD
>Gg_ Angiopoietin-like 4_XP_001232284.2
MSQSGEKEQGTEPSGSEKSAQDHTNRRFTKRHDHQQLPAPIAALHHVSHPTGGCSVPPGGTWLLSLMKA
QEKVNLLIPLHPKDNKTQSPKWKINPKKSFSHTNQSHNVSVEPALPHKLPEDCQQLFLAGQQSSGVFQVQP
SGSQPFKVYCDMTAEGGWTVIQRRTDGSVDFDQLWDAYKNGFGDLHGDFWLGLEKIHHLVQEGRYDLLIE
LEDWEGNSQEIQFEFSLGGESTAYTLNLLGPLSGELENAIGDFRQLPFSTRDRDHDLKADTNCAKHLSGGW
WFSTCGHANLNGKYFRSIPRQRHERKQGIFWKTWKGRYYPLKSTTMKIQPAALEAEP
>Ip_ Angiopoietin-like 4_in silico translated from EST CK419825
GNHSDARAIQLQLEAQNQRIDELVERIKQQQEKLDKQNIRIRALQSQIQMRKERLNPSADEVRTEQQDTATA
SNCHDVFLRGETTSGVYTLQPRDSLPLHVYCEMTSDGGWTVIQRRRDGSVDFDQLWNEYQNGFGNLDGE
FWLGLEKMYRLTKDEDFILKIQMTDWRDEHQSVQYRFRLNGEDKNYSLQILESPDGNLESSLSTESSSLPFS
TRDKDNDWEYDFNCAKHLSGGWWFSNCGRSN
>Mm_ Angiopoietin-like 3_NP_038941.1
MHTIKLFLFVVPLVIASRVDPDLSSFDSAPSEPKSRFAMLDDVKILANGLLQLGHGLKDFVHKTKGQINDIFQK
LNIFDQSFYDLSLRTNEIKEEEKELRRTTSTLQVKNEEVKNMSVELNSKLESLLEEKTALQHKVRALEEQLTN
LILSPAGAQEHPEVTSLKSFVEQQDNSIRELLQSVEEQYKQLSQQHMQIKEIEKQLRKTGIQEPSENSLSSKS
RAPRTTPPLQLNETENTEQDDLPADCSAVYNRGEHTSGVYTIKPRNSQGFNVYCDTQSGSPWTLIQHRKD
GSQDFNETWENYEKGFGRLDGEFWLGLEKIYAIVQQSNYILRLELQDWKDSKHYVEYSFHLGSHETNYTLH
VAEIAGNIPGALPEHTDLMFSTWNHRAKGQLYCPESYSGGWWWNDICGENNLNGKYNKPRTKSRPERRR
GIYWRPQSRKLYAIKSSKMMLQPTT
>Mm_ Angiopoietin-like 4_NP_065606.2
MRCAPTAGAALVLCAATAGLLSAQGRPAQPEPPRFASWDEMNLLAHGLLQLGHGLREHVERTRGQLGALE
RRMAACGNACQGPKGKDAPFKDSEDRVPEGQTPETLQSLQTQLKAQNSKIQQLFQKVAQQQRYLSKQNL
RIQNLQSQIDLLAPTHLDNGVDKTSRGKKLSKMTQLIGLTSNATHLHRPARDCQELFQEGERHSGLFQIQPL
GSPPFLVNCEMTSDGGWTVIQRRLNGSVDFNQSWEAYKDGFGDPQGEFWLGLEKMHSITGDRGSQLAVQ
LQDWDGNAKLLQFPIHLGGEDTAYSLQLTEPTANELGATNVSPNGLSLPFSTWDQDHDLRGDLNCAKSLSG
GWWFGTCSHSNLNGQYFHSIPRQRQERKKGIFWKTWKGRYYPLQATTLLIQPMEATAAS
>Ol_ Angiopoietin-like 3_ENSORLP00000013132
MKLFLLLLVVVSSTAVVFSGNSGRQQVPTLPPEAFITAPTPTEIKSRFAMLDDVRLLANGLLQLGQSLREFVH
KTKAQINDIFQKLNIFDRSFYQLSVVTSEIKEEEEELKKTTSFLKANNEEIRNLSLEINSKINNILQERTQLQKKV

GSLEERLKGLSQSMIPSDQLSEITTLKEVIDAQERTITSLLKSVKEQHDQLDNQNTKIKHLEEKLSFDSFQDTV
DKPVDPDQTASDIFEYLTANTTGLEINDLPVDCSDLFNKGEDNSGIYMIKPNQSEPFYVYCEIDSDGSLTVIQR
RLDGSVDFDESWDKYEKGFGDLEKDFWLGLQKIHSLTQQRPYILRIDLEDWKEEKHWAEYHFVVGSPSTGY
TLHVSNFSGDLQDAMTNLNGMKFSTKDRSNNDQRDSSCARNNTGGWWQSMSCESNLNGKYLWMRAKG
RSVRRKGVHWRPRTGPSYYFKKTKITLRPAITANKA
>Ol_ Angiopoietin-like 4_ ENSORLP00000006679
MKTSLGTLTLCMVALMATGFPFERKGASGGAAKEKRVQYAAWDDVNVIAHGLLQLGQGLKEHVDKTKVQM
RDIFAKLKVFNRTMSELGKESQRLREEEEVLRARAQGLEDREGQLLNVTAELREKAEEIQQERRTMSERMS
RLEERVDSMLQGAAASPDMSAGARNNSHDGSIQDMVESQSRRINDLMERVRLQQDKLDKQNVRIRTLQSQ
IQQPRQRSSFPGRSSTEDPLQGGAAEQQDTPVVQSNRTAVDCHDLFLRGQTTSGVYSVQPAGSEPFSVFC
EMTADGGWTVIQRRQDGSVDFDQLWAAYEKGFGSLEGEFWLGLQNIHSIVQDNSYVLSIQLSDWGEDFVS
VHLPIQLGGKQTNYSLLVQEPVAANSLQSSLGSDSAAAGLSFSTPDRDNDLKVDISCAKLLSGGWWFSNCG
RSNLNGRYFQSPPPKQRHQRKQGIFWKTWRGRYYPLKSSVMMIAP
>Ss_ Angiopoietin-like 3_NP_001003926.1
MYTIKLFLLIAPLVISSRIDQDSSSLDSVSPEPKSRFAMLDDVKILANGLLQLGHGLKDFVHKTKGQINDIFQKL
NIFDQSFYDLSLQTNEIKEEEKELRRTTFKLQVKNEEVKNMSLDLNSKVESLLEEKILLQHKVRYLEDQLTNLI
KNQPEIQEHPDITSLKTFVEQQDNSIRDLLQTVEEQYRQLNQQHSQIKEIENQLRRTGTQESTENASKPRVP
RTTPSLHLNETRNVEHNDIPADCTVIYNRGDQTSGIYSIRASNSQVFNVYCDVKSGSSWTLIQHRIDGSQNF
NETWENYRYGFGRLDGEFWLGLEKIYSIVKQSNYILRIELEDWNDNEYYIEYSFHLGDHETNYTLHLVEIAGN
VPNALPEHEDLMFSTWDHKAKGHVNCPESYSGGWWCHDVCGENNLNGIYKKPKAKIKPERRGICWKSQN
GRLYSIKSTKMLIHPIDSESSELTKAIA
>Ss_ Angiopoietin-like 4_NP_001033733.1
MRSAPTARAALVLCAATAGLLSAQGSPEPPEAPRFASWDEVNVLAHGLLQLGRGLREHVERTRGQLGALE
RRLSACGAACKDPEGSAVPPLTAGNLVPSQSDAAPETLHSLQTQLKAQNSKIQQLFQKVAQQQRHLEKQHL
RIQNLQGQLDHLAPMHLGHGVAKAARRKRLPKMTQPAGPAHNISRLHRLPRDCQELFEEGERQSGLFQIQP
QGSPPFLVNCKMTSDGGWTVIQRRQDGSVDFNQPWEAYKDGFGDPKGEFWLGLEKVHRIMGDRGSRLA
VQLQDWEGNAESLQFPVHLGGEDTAYSLQLTAPVASKLGATIDTPSGLSLPFSTWDQDHDLRGDKNCAKIP
SGGWWFGTCSHSNLNGQYFHSIPRQREQRKKGIFWKTWRGRYYPLQATTMLIQPTVAEVAS
>Tr_ Angiopoietin-like 3a_ENSTRUP00000047391
MKLLYLLLLASCTAATPLESSSREKYTTLPSDVFTTATMPPEAKSRFALLDDVRLLANGLLQLGQSLREFVHK
TKGQINDIFQKLNIFDRSFYQLSVVTSEIKEEEAELKKTTNYLKANNEEIKNLSLEINSKINSILQERAQLQSKVG
NLEEKMQGLSQSMVPLDHVNEITTLKEVIETQEKTIGILLNAVREQHDQLNNQKIKITNLEDKISYDNYQDTVD
KAKYPDPDISDLFEYLAGNSSLDTNELATDCSELFDKGETNSGIYVIKPNQSEPFYVYCEMGSDGGSTVVQR
RVDGSVEFNQSWNKYELGFGDLQNDFWLGLEKIYSLTQQGDYILRIDLEDWKEERHWAEYQFSLEGPSKD
YIIQVTSFSGDLPDALANSTGMRFSTKDRNTDDNQNSNCNRSYTGGWWVNACGETHLNGRYQWLRAKGR
APRRRGIHWRPAAGPSFYLKMTKMTLLPVQHTNQH
>Tr_ Angiopoietin-like 3b_ENSTRUP00000004896
AHVLLLVLLSAGVPALCEPKEQPGVQPVAPTQAPRSRFAALDDVRLLGNGLLQLGQSLREFVQKTKGQINDI
FQKLSIFDRSFNQLSVLTSEIKEEEEELKKTTVVLKASNDEIKGLSVQIVSKVDSILQEKSNLHDKLEGLEEKLS
SLSNGLVPRQQAAEINSLREVIHSQETSIRELLRAVTDQSDQLNLQRMKIKTLEEKLSRKPQETIEKIPEVFSS
EMPMLSAHQPPHLTSTSELMRDLPSDCSQLFDRGARVSSVYSIQPHGSEPFMVFCDMSKGHGETVIQRRM
DGLINFDQTWETYENGFGALQEEFWLGLRNIRSLLVRGNSVLHVQLEDWKQGRHSSEYTFYLHGPEEDYVI
DLRLLSGDLPDPMGNLTGMAFSTKDRDSDQQRDSDCAHGYTGGWWFNACGDAFLNGKYFQMRPKGRTE
RRKGIQWRSGPKAFTSLMSTQISVRQMAPPSSVSSTSS
>Tr_ Angiopoietin-like 4a_ENSTRUP00000023243
MKTTLATLTLCLVVLMATGFPFDRKGGSAAGGGSKEKRVQYAAWDDVNVIAHGLLQLGQGLKEHVDKTKV
QMRDVSTKLKVFNRTMTDLAKESQKLRVEGEALKGRARELEDREGQLLNVTAELREKAEEMQLERRAMSE
RMSRLEERVDSLLQGGGVLPDLEAGAKNSSDARHIQVMLENQNRRIDDLLERIRLQQEKLDKQNVRIRTLQS
LLRSHGAFLDFPQVVESRNGDASVEQSDSPIETVSDCHELFLRGETTSGVYTIQPVNAEPFKVFCEMTADG
GWTVIQRRQDGSLDFDQLWEAYVKGFGSLTGESWLGLEKIHSIAKDGGYILNIQLTDWNGDVASVKLPFSL
GGGESKYSLQVRKDGPFSPLERSLGADVLHGLPFSTRDQDNDQKNDTNCAKHLSGGWWFSSCGHSNLN
GRYFQSPPPKHRHQRKQGIFWKSWRGRYYPLKNTVMLIAPVSVQSKS
>Tr_ Angiopoietin-like 4b_ENSTRUP00000018643

MKDRVLVVSRVINNYKSLPSRFKVAVEPLDKYASWDDVNVVSHGLLQLGQGLKEHVDKTKAQTRDVNAKLK
SLDAAVEEVERRQRKQDEALRAGSKEAEDREKLLAALAEEVEEVKKQSKNINSKVDKLEEKLEDGGHLGVS
RGCLQKMVAAQNRRIDQLVEKLEQQQDKLDKQSLHLQMLQTKVSRGSSRATTHPSAAGNRPHLSSCSTRF
HTRVSGGVRDCHHLYVRGQRASGVYTIQPEGSEPFAVFCDMTSEGGWTVIQKRYDGAQNFNQLWEGYKR
GFGSLDGEFWLGLEKIRSVSKQGPYQLQVELSDGAGQQLPVARYLFQLDGEEKKFALHLEDEAPSPRTSTG
SSGIPFSTADRDNDLSEDVSCAKLLSGGWWFSSCGDWNLNGRFPRRPSGPSRKQTRKMFWTSKGQRHS
VRTTLLKIAPTTMKLRS
>Tn_ Angiopoietin-like 3_ENSTNIP00000010672
TPVLLLALLFVGVPALCDSKEELFLQTTAPTQAPRSRFAALDEVRLLANGLLHLGQSMREFVQKTRGQISDIF
QKLNIFDRSFYQLSVLTSEIKEEEEELKKTTVVLKASNDEIKDLSAQISFKVDSILQEKSDLQDKLEGLEEKLSS
MSKSAPLRYQAAEINNSIHTQEVIHSQDNSIRELLRAVRHQSHQLNLHRVKIKSLEEKLTTGKKPQETVERISE
VSSAETPTLSPYQASHSASTSELMNLPSDCSQLFESGVRISSVYAIRPHSSEPFVVFCDMSEDHGETVIQRR
MGGLVNFDQTWETYENGFGDLQGEFWLGLSSIRSLLARGNTVLRVQLEDWKQGSHLSEYNFYLSGPEEDY
TINLRLLSGDTPDPMGNLTGMAFSTKDRNSDQQQDSSCAYGYTGGWWFNACGDAHLNGKYFQLRPKGIQ
WRSGPKAFTSLKSTKISIRHMAPPSSVSSP
>Tn_ Angiopoietin-like 4a_ENSTNIP00000003092
MKRTLATLTFCLLVLVATGFPFDKKGGSAAGGASKEKRVQYAAWDDVNVIAHGLLQLGQGLKEHVDRTKVQ
MRDISTKLKLFNRTVTDLGKESQKLRAEGEAAKSRARELEDREGQLLNITAELREKAEEMQLERRAMSARM
SRLEEMLPGAEAGAGNGSDARHIQVMLENQNRRIDDLLERIRLQQEKLDKQNARIRTLQNQVGGTKFRSVE
TRPRVEPDACKQAAAKKTPILLLISGNAEMASDCHELFLRGEATSGVYTIQPVNAQPFKVFCEMTASGGWTV
IQRRQDGSVDFDQLWEAYLRGFGSLNGEFWLGLEKIHSISKDGGHILNIQLSDWNGDVASVTLPFSLGGEET
QFSLQVRKDGPLSTLERSLGADAHGGLPFSTRDQDNDRKNDTSCAKHLSAGGWWFSSCGHSNLNGRYFQ
SPPPKHRHQRKQGIFWKSWRGRYYPLKKSVMMVAPASVQSKS
>Tn_ Angiopoietin-like 4_ENSTNIP00000013099
MKTPQLLVLLSSTLVGVSTAFPAHRSPDPDQDPDQDQYASWDEVNVVAHGLLQLGQGLKEHVDKTKAQTR
DINTRLKLLDATVVEVERRWREQEEALRARSSQVEEREKLLAEVAQEVRGKVEEVKKQSQNMDQLEKKGG
SDPSWGGWCLQKVLAAQNSRIDPLVEKMEQQEDKLDKQSLRLQRLESKQNTASASTLPRQVSHRRAQRR
RDGKPREEEPRSAGGHVCALLSGRARDCQHLYAAGQRASGVYTIQPDGSHPLDVFCDMTSEGGWTVIQR
RHDGSQNFNQPWERYKRGFGSLSGEFWLGLEKIRSVSKQGPYQLRVELSNGAGQQLPVARYGFHLDGED
KKFALRLEDETASPPATAGGSGIPFSTADRDNDLAVDVNCAELLSGGWWFSSCGDWNLNGRRPSAPSREQ
PRKPEAFRTSQGRRRSVKTTLLKIAPTGTGV
>Xt_ Angiopoietin-like 3_ENSXETP00000023732
MNLVLIFILPLVLSATEKDDSAYDSLSTDSKSRFAMLDDVRILANGLLQLGHGLKDFVHKTKGQINEIFQKLNIF
DKSVTDLSEQTNEIREKEEELKDTTSKLQENNEELKNISRKINSQVENLLQDKIHLQAKVGSLEEKLFQMTQG
TTEGQEIKEISSLKNFVEQQDVNIRHLLKVVQEQHMQLDHQNVQIKDLEDKLSKADLQESVKSVLAVRRSRT
GFLNLSNSTDGMVEQNDSRDCNDIYNRGERSSGIYTIRPNGSTAFDVYCEITSESANTVIQRRTDGSVDFNQ
TWETYLNGFGELTGEFWLGLEKIHAISQQADYILHIELQDWKENWRFVEYMFTLGNQDTSYALQLTQVSGNI
PSALPEQREILFSTSDQNSGDLKCPAETFSGGWWNTACSGTNLNGKYIKQRPRTKLDRRRGQGIYWKSEK
GRLYSLKSTKIMLYRTDLDSFE
>Xt_ Angiopoietin-like 4_ENSXETP00000047381
MKLLLASITVSLLVLSLLVLGGESWGFSSEKKVQYASWDEVNVLAHGLLQLGHGLKEHVDKTKGQLKEISGK
LVQHNVSLLELSRQASEVRESGEALKGRLQELEDKDKQLYDVSQGLKGKVQEISKDRQLLEHRLQNMEAKI
QLLEPSKRQNRSEKEDLLSIQTLMEIQSKRIDELLEKIKLQQYKLDKQNLQIKSLQNTVSLTFIANTKHILQIQSN
RLETQTWKMNLKKTVEDEVFPSDCHQIFLEGKKSSGIFSIQPSGAQPFEVYCEMTADAGWTVTQRRTDGSV
DFDRLWDAYTDGFGNLNGEFWLGLEKMHQITQQGQYLIHIDLQDWENNVQHMEAKFLLAGSNEAYALQLL
GPVTGELENALSDFQQLQFSTRDRDQDKKSDFNCAKHLSGGWWFSSCGHSNLNGKYFLSVPRARHERKQ
GIFWKTWKGRYYPLKSTSIKIRPVDTDLTV

Table S2: Primer sequences used in this study

Transcription start site mapping:

Name	Forward	Reverse
Angptl4 5'RACE F	GAGAGTGCTGCGACAGCTAA	
Angptl4 5'RACE R2		GCAGCTCCTCTCCTCTCCAT
Angptl4 5'RACE R1		ATCATCCCACGCTGCATATT

Promoter reporter analysis:

Name	Forward	Reverse	Cloning Method
cfos	TAATCTCGAGGTTTAAACCCAGTGACGTAGGAAGTCCA	TAATGGATCCCGAAGTTTGGGGAAAGCCCG	BamHI/XhoI
Dr -5.2kb	TAATCTCGAGGCGATATTTTACTGTTCTGC	TAATGGATCCCTTGCTTGTTTTGTGTCTGTAG	BamHI/XhoI
Dr -3.5kb	TAATCTCGAGCTTCTGTGATGCAGATGTTGTG	TAATGGATCCCTTGCTTGTTTTGTGTCTGTAG	BamHI/XhoI
Dr -1kb	TAATCTCGAGTCCTGGCTTCTGCTCCACTTAC	TAATGGATCCCTTGCTTGTTTTGTGTCTGTAG	BamHI/XhoI

Truncation mapping:

Name	Forward	Reverse	Cloning Method
Dr in3	GGGGACAAGTTTGTACAAAAAAGCAGGCTGTATAAGTCAGTTTATCATT	GGGGACCACTTTGTACAAGAAAGCTGGGTACTGAAAGACACAAACACA	Gateway Recombination
Dr in3.1	GGGGACAAGTTTGTACAAAAAAGCAGGCTGTATAAGTCAGTTTATCATT	GGGGACCACTTTGTACAAGAAAGCTGGGTGAATCAGAGCCTTCATTACTTGA	Gateway Recombination
Dr in3.2	GGGGACAAGTTTGTACAAAAAAGCAGGCTGTCAGTTAATGTAGGGCATCC	GGGGACCACTTTGTACAAGAAAGCTGGGTACTGAAAGACACAAACACA	Gateway Recombination
Dr in3.3	GGGGACAAGTTTGTACAAAAAAGCAGGCTGTCAGTTAATGTAGGGCATCC	GGGGACCACTTTGTACAAGAAAGCTGGGTGGTGTGTTTGGTTAAACATCTGC	Gateway Recombination
Dr in3.4	GGGGACAAGTTTGTACAAAAAAGCAGGCTCCTTGTAGGCTGTTGGAAATAC	GGGGACCACTTTGTACAAGAAAGCTGGGTACTGAAAGACACAAACACA	Gateway Recombination
Dr in3.5	GGGGACAAGTTTGTACAAAAAAGCAGGCTCCTTGTAGGCTGTTGGAAATAC	GGGGACCACTTTGTACAAGAAAGCTGGGTGCAAATGACCTTTGCCCAAG	Gateway Recombination
Dr in3.6	GGGGACAAGTTTGTACAAAAAAGCAGGCTCCTTGTAGGCTGTTGGAAATAC	GGGGACCACTTTGTACAAGAAAGCTGGGTCTGTACCCAAACAAAGATGT	Gateway Recombination
Dr in3.7	GGGGACAAGTTTGTACAAAAAAGCAGGCTCCTTGTAGGCTGTTGGAAATAC	GGGGACCACTTTGTACAAGAAAGCTGGGTTTGAACATGTCACCCAAGAACC	Gateway Recombination
Dr in3.8	GGGGACAAGTTTGTACAAAAAAGCAGGCTGGTCCAGTGTTTGAGATAAG	GGGGACCACTTTGTACAAGAAAGCTGGGTACTGAAAGACACAAACACA	Gateway Recombination
Dr in3.9	GGGGACAAGTTTGTACAAAAAAGCAGGCTACTCCATGCTGGCTTCTGGG	GGGGACCACTTTGTACAAGAAAGCTGGGTACTGAAAGACACAAACACA	Gateway Recombination
Dr in3.10	GGGGACAAGTTTGTACAAAAAAGCAGGCTGTTTAATAACAACACTTCAG	GGGGACCACTTTGTACAAGAAAGCTGGGTACTGAAAGACACAAACACA	Gateway Recombination
Dr in3.11	GGGGACAAGTTTGTACAAAAAAGCAGGCTACTCCATGCTGGCTTCTGGG	GGGGACCACTTTGTACAAGAAAGCTGGGTGCAAATGACCTTTGCCCAAG	Gateway Recombination
Dr in3.12	GGGGACAAGTTTGTACAAAAAAGCAGGCTACTCCATGCTGGCTTCTGGG	GGGGACCACTTTGTACAAGAAAGCTGGGTAACGCTCACACGTCCAGGAT	Gateway Recombination
Dr in3.13	GGGGACAAGTTTGTACAAAAAAGCAGGCTGTCAGTTAATGTAGGGCATCC	GGGGACCACTTTGTACAAGAAAGCTGGGTGAGGAAAGTCTCTGCTTGGG	Gateway Recombination
Dr in3.14	GGGGACAAGTTTGTACAAAAAAGCAGGCTGTCAGTTAATGTAGGGCATCC	GGGGACCACTTTGTACAAGAAAGCTGGGTGGGAGGAAAGAATGCTTGTA	Gateway Recombination
Dr in3.15	GGGGACAAGTTTGTACAAAAAAGCAGGCTGTGGAGAGCCGGACTAAAGG	GGGGACCACTTTGTACAAGAAAGCTGGGTGGTGTGTTTGGTTAAACATCTGC	Gateway Recombination
Dr in3.16	GGGGACAAGTTTGTACAAAAAAGCAGGCTACCAGCAGCCACGGCATTCC	GGGGACCACTTTGTACAAGAAAGCTGGGTGGTGTGTTTGGTTAAACATCTGC	Gateway Recombination
Dr in3.17	GGGGACAAGTTTGTACAAAAAAGCAGGCTACCAGCAGCCACGGCATTCC	GGGGACCACTTTGTACAAGAAAGCTGGGTGAGGAAAGTCTCTGCTTGGG	Gateway Recombination
in3.4 ds-inv	TAATAGATCTGGCCGGCCCCTTGTAGGCTGTTGGAAAT	TAATAGATCTGGCGCGCCTGAAAGACACAAACACATG	BglII

Site-directed mutagenesis:

Name	Forward	Reverse	Cloning Method
Sub1	ACCCTGGAGTTCAAAACAAAAGATCGATGTGCTTCTGGGTTCTTGGGTGA	TCACCCAAGAACCCAGAAGCACATCGATCTTTTGTTTTGAACTCCAGGGT	Circular PCR/DpnI/ClaI
Sub2	CAAAACAAACTCCATGCTGGAAATCGATGGCTTGGGTGACATGTTCAAGG	CCTTGAACATGTCACCCAAGCCATCGATTTCCAGCATGGAGTTTGTTTTG	Circular PCR/DpnI/ClaI
Sub3	GGTTTGTCACTGTTCAAGGTCCAGTGTTTG	GTGACAAACCGAACCCAGAAGCCAGCATGG	Linker mediated PCR
Sub4	TCTGGGTTCTTGGGTGACATTAATCGATGACAGTGTTTGAGATAAGGATT	AATCCTTATCTCAAACACTGTCATCGATTAATGTCACCCAAGAACCCAGA	Circular PCR/DpnI/ClaI
Sub5	TGGGTGACATGTTCAAGGTCACATCGATTCGATAAGGATTTAGTACACCA	TGGTGTACTAAATCCTTATCGAATCGATGTGACCTTGAACATGTCACCCA	Circular PCR/DpnI/ClaI
Sub6	TGTTCAAGGTCCAGTGTTTGCAATCGATGGTTAGTACACCATTAAACAAT	ATTGTTTAATGGTGTACTAACCATCGATTGCAAACACTGGACCTTGAACA	Circular PCR/DpnI/ClaI
Sub7	CCAGTGTTTGAGATAAGGATGCATCGATAAATTAAACAATAAGATAAACA	TGTTTATCTTATTGTTTAATTTATCGATGCATCCTTATCTCAAACACTGG	Circular PCR/DpnI/ClaI
Sub8	GATAAGGATTTAGTACACCAGAATCGATGCAGATAAACACTTATCCTGGA	TCCAGGATAAGTGTTTATCTGCATCGATTCTGGTGTACTAAATCCTTATC	Circular PCR/DpnI/ClaI
Sub9	TAGTACACCATTAAACAATACAATCGATGGTTATCCTGGACGTGTGAGCG	CGCTCACACGTCCAGGATAACCATCGATTGTATTGTTTAATGGTGTACTA	Circular PCR/DpnI/ClaI
Sub10	TAAACAATAAGATAAACACTGCATCGATGAGTGTGAGCGTTTTAAATACT	AGTATTTAAAACGCTCACACTCATCGATGCAGTGTTTATCTTATTGTTTA	Circular PCR/DpnI/ClaI
Sub11	ATAAACACTTATCCTGGACGGTATCGATGGTTAAATACTTTGGCAACTTT	AAAGTTGCCAAAGTATTTAACCATCGATACCGTCCAGGATAAGTGTTTAT	Circular PCR/DpnI/ClaI
Sub12	ATCCTGGACGTGTGAGCGTTGCATCGATGGTGGCAACTTTAACATCTTTG	CAAAGATGTTAAAGTTGCCACCATCGATGCAACGCTCACACGTCCAGGAT	Circular PCR/DpnI/ClaI
Sub13	TGTGAGCGTTTTAAATACTTGTATCGATGGAACATCTTTGTTTGGGTACA	TGTACCCAAACAAAGATGTTCCATCGATACAAGTATTTAAAACGCTCACA	Circular PCR/DpnI/ClaI
Sub14	TTAAATACTTTGGCAACTTTCCATCGATGTTTTGGGTACAGCCTTGGGCA	TGCCCAAGGCTGTACCCAAAACATCGATGGAAAGTTGCCAAAGTATTTAA	Circular PCR/DpnI/ClaI
Sub15	TGGCAACTTTAACATCTTTGGCATCGATACGCCTTGGGCAAAGGTCATTT	AAATGACCTTTGCCCAAGGCGTATCGATGCCAAAGATGTTAAAGTTGCCA	Circular PCR/DpnI/ClaI
Sub16	AACATCTTTGTTTGGGTACATAATCGATACAAGGTCATTTGCAGATGCTT	AAGCATCTGCAAATGACCTTGTATCGATTATGTACCCAAACAAAGATGTT	Circular PCR/DpnI/ClaI
Sub17	TTTGGGTACAGCCTTGGGCACCATCGATGGGCAGATGCTTGAACATGTGT	ACACATGTTCAAGCATCTGCCCATCGATGGTGCCCAAGGCTGTACCCAAA	Circular PCR/DpnI/ClaI

Angptl4 intron 3 cloning for sequencing from multiple species:

Name	Forward	Reverse	Cloning Method
Danio_angptl4_ex3-4	CTGAACGCGAGCACTGAACAGCGAG	AACAACTCATGACAGTCAGAAGCCA	Topo TA
Cyp anpgptl4 ex3-4	CTGAACGCGAGCTCTGAACAGCGAG	AACAGCTCATGACAGTCAGAAGCCA	Topo TA
Ip angptl4 ex3-4	CCCGAGTGCTGATGAAGTTC	TCTCTCCACGTAGGAACACA	Topo TA
Danio in3.2 F	GTCAGTTAATGTAGGGCATCC	AACAACTCATGACAGTCAGAAGCCA	Topo TA

Intron 3 for reporter assay from multiple species:

Name	Forward	Reverse	Cloning Method
Cyprinoforme_in3.2	GGGGACAAGTTTGTACAAAAAAGCAGGCTGTCAGTTAATGTAGGGCATCC	GGGGACCACTTTGTACAAGAAAGCTGGGTACTGAAAGACACAAACACA	Gateway Recompbination
Ol in3.2	GGGGACAAGTTTGTACAAAAAAGCAGGCTCTCTGCTGATGAAGGAATCC	GGGGACCACTTTGTACAAGAAAGCTGGGTCTGAAAGACATTTTAGGGC	Gateway Recompbination
Ip in3.2	GGGGACAAGTTTGTACAAAAAAGCAGGCTTCAGTGACTTACACAGCGAG	GGGGACCACTTTGTACAAGAAAGCTGGGTCTGCAAAATAACATTAAACAC	Gateway Recompbination

qRT-PCR assays:

Gene	Forward	Reverse
angptl4	CGAGCGCATCAAGCAACA	TCGCTCGTTTTTCATCGTAATCT
egfp	GAAGAAGTCGTGCTGCTTCA	CCTGAAGTTCATCTGCACCA
18S	CACTTGTCCCTCTAAGAAGTTGCA	GGTTGATTCCGATAACGAACGA
tdT	CACCATCGTGGAACAGTACG	GGCCATGTTGTTGTCCTC
rpl32	CCCTCACCAAACCTAAGATCGT	CTCCAGTTTGCCCTGATCTTG

Dnase I hypersensitivity:

Region	Forward	Reverse
TATA box	CTGAGCAGACTCGCACACTC	CTCATGCTGTTGAAAGGCTAACT
in3.3	GATGTCATCAGAGGGGTGCT	CTCAGTCCCTCAGAGCTCTACAA
in3.4	CTGGACGTGTGAGCGTTTTA	GACACAAACACATGTTCAAGCA

Table 3.S2: Primer sequences used in this study.

Abbreviated name used in paper	Full allele name[a]	Tissue specificity[b]	Propagated[c]
Tg(in3.4-Mmu.Fos:GFP)	*Tg(in3.4angptl4-Mmu.Fos:GFP)nc2*	intestine	Yes
	Tg(in3.4angptl4-Mmu.Fos:GFP)nc3	intestine	Yes
	Tg(in3.4angptl4-Mmu.Fos:GFP)nc4	intestine	No
Tg(-5.2angptl4:GFP)	*Tg(-5.2angptl4:GFP)nc5*	liver	No
	Tg(-5.2angptl4:GFP)nc6	liver	No
Tg(-3.5angptl4:GFP)	*Tg(-3.5kbangptl4:GFP)nc7*	liver	No
	Tg(-3.5angptl4:GFP)nc8	liver	No
Tg(-1angptl4:GFP)	*Tg(-1angptl4:GFP)nc9*	liver	No
	Tg(-1angptl4:GFP)nc10	liver	No
Tg(in3-Mmu.Fos:GFP)	*Tg(in3angptl4-Mmu.Fos:GFP)nc11*	liver/islet/intestine	Yes
	Tg(in3angptl4-Mmu.Fos:GFP)nc12	liver/islet/intestine	No
Tg(in3.1-Mmu.Fos:GFP)	*Tg(in3.1angptl4-Mmu.Fos:GFP)nc13*	liver	No
	Tg(in3.1angptl4-Mmu.Fos:GFP)nc14	liver	No
Tg(in3.2-Mmu.Fos:GFP)	*Tg(in3.2angptl4-Mmu.Fos:GFP)nc15*	islet/intestine	Yes
	Tg(in3.2angptl4-Mmu.Fos:GFP)nc16	islet/intestine	No
Tg(in3.2-Mmu.Fos:tdT)	*Tg(in3.2angptl4-Mmu.Fos:TdTomato)nc17*	islet/intestine	Yes
	Tg(in3.2angptl4-Mmu.Fos:TdTomato)nc18	islet/intestine	Yes
Tg(in3.3-Mmu.Fos:GFP)	*Tg(in3.3angptl4-Mmu.Fos:GFP)nc19*	islet	Yes
	Tg(in3.3angptl4-Mmu.Fos:GFP)nc20	islet	No
Tg(-1angptl4:GFP:in3.4inv)	*Tg(-1angptl4:GFP:in3.4inv)nc21*	liver/intestine	No
	Tg(-1angptl4:GFP:in3.4inv)nc22	liver/intestine	No
Tg(lpin3.2-Mmu.Fos:GFP)	*Tg(lpu.in3.2angptl4-Mmu.Fos:GFP)nc23*	islet	Yes

[a] Full allele names provided in compliance with the ZFIN Zebrafish Nomenclature Guidelines.

[a] Tissue specificity of fluorescent protein expression. Note that all lines also display muscle expression, however the intensity of muscle expression varied depending on the construct.

[c] Some transgenic lines were propagated for future use, while others were terminated.

Table 3.S3: Allele designations for stable lines created in this study.

CHAPTER 4

Towards Identification of Transcription Factors Regulating Intestinal Expression of *Angptl4*

4.1 Overview

The regulatory potential of *cis*-regulatory modules (CRMs) is facilitated by protein factors that bind DNA in *trans* to specify a genomic locus for transcriptional activation or repression (henceforth transcription factors). A major challenge is to identify the corresponding transcription factors that bind a defined CRM. I previously utilized the unique features of the zebrafish model to elucidate and characterize the *in vivo* activity of multiple CRMs that confer tissue-specificity and microbial control of *angptl4* transcription, a circulating inhibitor of lipoprotein lipase (LPL). This work elucidated a 40 base pair region (termed in3.4-CR) within the third intron of zebrafish *angptl4* that is required for intestinal expression. To discover factors that regulate *angptl4* intestinal transcription, I first established an electrophoretic mobility shift assay (EMSA) using nuclear extracts harvested from zebrafish intestinal epithelial cells. I used this assay to reveal that multiple complexes can assemble on an in3.4-CR double-stranded DNA probe and used mutant probes to localize required binding sites. The same bases required for *in vivo* reporter experiments were also required for *in vitro* shift activity. I used conditions established in this EMSA assay to try to identify the corresponding transcription factors by DNA affinity chromatography followed by mass spectrometry. I identified multiple protein factors that appear to preferentially bind the in3.4-CR regions compared to a control probe, however these candidate factors were not predicted to be

sequence-specific transcription factors. These results provide a foundation for future efforts aimed at the unbiased discovery of factors regulating transcription in the zebrafish intestinal epithelium.

4.2 Introduction

Pioneering work on lambda phage during the 1960s led by Francois Jacob and Jacques Monod [72] predicted the discovery of factors that bind in *trans* to non-genic DNA regions and function to modulate expression of protein-coding genes [245,246]. The biochemical isolation of the lac and lambda repressors and their subsequent molecular understanding [247] revealed that these *trans* acting factors are proteins that have a binding specificity for particular DNA sequences. This early work in phage, and also *E. coli* [248,249], was extended to eukaryotic cells by a number of laboratories in the 1970's and 1980's [250-254]. It is now firmly accepted that all known life-forms selectively regulate RNA transcription in part through the sequence-specific binding of protein factors to regulatory DNA. How then does specificity arise?

As of the writing of this thesis, there are 1,764 transcription factor structures deposited in the Protein Data Bank (PDB). A number of insights have been inferred from this wealth of structural information [255,256]. Notably, it appears that transcription factors are typically modular in their structure, often consisting of a trans-activation domain and DNA-binding domain (DBD) [256,257]; modular in that the DBD can alone bind specifically to its cognate DNA sequence [258]. The DBD typically binds 6-10 base pair DNA sequences (or motifs) and multiple families of domain configurations (fingers, zippers, helices, homeoboxes, etc) have evolved to recognize DNA. Binding site recognition proceeds in part through specific interactions with bases and nonspecific interactions with the negatively charged sugar/phosphate backbone. It is believed that

specificity is largely conferred by “sequence-dependent projections of chemical groups from the bases into the major and minor grooves” of the DNA polymer [259,260]. Chemical interactions including direct (and indirect via a water molecule) hydrogen bonding and van der Waals (VdW) interactions that occur between amino acid side chains and available chemical groups of bases projected at the protein-DNA interface. Each dinucleotide base pair (TA, AT, GC, CG) harbors a distinctive capacity to contribute to H-bonding or VdW interactions in both the major and minor groove, which is determined by the chemical group display array of that dinucleotide. Therefore, varying the pattern of dinucleotides can generate distinct binding pockets. Notably, this feature suggests that abstracting binding sites to linear text can be used to predict the transcription factors capable of binding a given DNA sequence. More accurately, the binding preferences of transcription factors are often degenerate and can be better modeled through the likelihood that a certain base is present at a given position within the binding window of the transcription factor’s DNA binding domain. These likelihoods (or motifs) are commonly represented as position weight matrices (PWMs) and can be used to ask if a similar motif is found within a given DNA text string. It is believed that recognition is further influenced by the three-dimensional structure, conformation, and deformability inherent to the chromatin DNA region. Indeed, accounting for local DNA topography has been shown to be a better predictor of transcription factor recognition [261], and *cis*-regulatory regions in general [262], than primary sequence alone. It must be noted that transcription factors can often bind multiple apparently un-related recognition sites and this discrepancy is not fully understood.

Vertebrates are composed of hundreds to thousands of different cell types, each containing a shared genome consisting of gigabases of DNA encoding tens of thousands of genes, and an infinite variety of environmental circumstances to encounter. As an example, a recent survey suggested that the human genome encodes

approximately 1,700-1,900 sequence-specific transcription factors [263]. It is difficult to imagine how development and homeostasis of such a complicated organism is maintained by a limited set of factors if each functions alone. One would imagine that the addition of an adjacent recognition site for another transcription factor would dramatically increase cooperative specificity and allow for combinatorial interactions to explain the observed discrepancy between regulatory needs and regulatory factors. Indeed, most tissue-specific *cis*-regulatory modules tend to harbor functional binding sites for 4-6 transcriptional regulators [222], and binding site clustering [264] and combinatorial binding [265] have been useful methods for predicting cis-regulatory activity [266]. A particularly elegant example of cooperative binding was revealed in the structure of the Interferon-ß enhanceosome in which 8 transcription factors cooperatively assemble on a 50 bp enhancer where at least one protein-DNA contact occurs at every base position [84]. The generality and logic of cooperative or combinatorial interactions is not fully understood, but likely plays an important role in reaching thermodynamic thresholds required for activation or repressive activities [134,267].

There are numerous DNA-based strategies one can employ towards the identification of candidate DNA-binding proteins [268]. Recognition sites for many transcription factors have been defined through experimental studies and high-quality PWMs have been deposited into public databases such as JASPAR and TRANSFAC. One approach towards transcription factor discovery is to query these databases for motifs present within the DNA regulatory module [202]. Coupling this search with evolutionary conservation can improve the predictive capacity of this strategy [269]. Current PWM databases include information from a very limited number of genomes and are incomplete for those genomes that are included. Further, many transcription factors bind a range of DNA sequences or have minimal specificity. Therefore, reliance on informatics approaches can limit the discovery power and can in principle lead to

systemic bias. A second approach uses biochemical purification of sequence specific DNA binding factors from cellular extracts by DNA chromatography and has historically been a useful method for identification of transcriptional regulators [270-272]. However, successful enrichment requires a binding activity assay and an abundant source of nuclear extract. Recent advances in mass spectrometry now allow efficient peptide identification in low abundance samples and complex mixtures [273,274] and application of these technologies to discover factors capable of binding DNA using relatively simple pull-downs are now possible [275,276]. The zebrafish field has traditionally relied on the predictive capacity of transcription factor binding site (TFBS) databases, in which most vertebrate TFBS motifs were defined in mammalian systems.

The intestinal microbiota affects vertebrate energy balance and fat storage through impacting the capacity for dietary energy harvest [10]. Previous work demonstrated that conventionally-raised mice (CONVR) harboring a healthy microbiota have increased fat storage when compared to germ-free (GF) mice lacking a microbiota [8]. This increase in fat storage in CONVR animals was due in part to decreased expression of Angiopoietin-like 4 (Angptl4) in the intestine of CONVR mice. Angptl4 is a key regulator of lipid deposition and functions by directly inhibiting Lipoprotein lipase (LPL) mediated hydrolysis of circulating triacylglycerides into free fatty acids and glycerol. In my previous work, I used *in vivo* reporter assays in the zebrafish to elucidate a CRM (termed in3.4) in the third intron of zebrafish *angptl4* that conferred intestine-specific expression. I used comparative sequence analysis from 12 fish species, functional mapping, and site-directed mutagenesis to define the minimal set of regulatory sequences required for intestinal activity of the *angptl4* intestinal CRM. Finally, I showed that this intestinal module also responds to the microbiota similar to the endogenous gene. These results provided a mechanism by which the microbiota might differentially regulate *angptl4* expression and peripheral fat storage by suppressing the activity of an

intestine-specific transcriptional enhancer. The transcription factors mediating intestinal expression and microbial suppression are unknown.

In the following chapter, I use a computational approach to predict regulatory motifs present in the in3.4 module and the candidate factors controlling intestinal expression of *angptl4*. In parallel, I developed a biochemical assay and DNA affinity pull-down strategy to discover regulators of *angptl4* transcription. The short-term goal was to establish a relatively simple method to identify protein factors capable of specifically binding the in3.4 regulatory region of *angptl4*. The longer-term rationale was to explore the utility of this strategy as a platform for medium-throughput unbiased discovery of transcription factors mediating microbial response in the intestine. Ultimately, I was not successful in decisively identifying a likely candidate regulator of intestinal *angptl4* transcription using this approach. However, this work constitutes an important step forward in assay development and provides the foundation toward using DNA affinity chromatography and mass spectrometry to identify transcription factors active in the zebrafish intestine.

4.3 Results

4.3.1 Computational prediction of candidate factors

I queried the JASPAR and TRANSFAC databases for transcription factor motifs present in the zebrafish in3.3 (islet) and in3.4 (intestine) modules (see Chapter 3, Figure 3.7) to discover candidate transcription factors that may regulate intestinal expression of *angptl4*. Because the in3.3 module does not drive expression in the intestine I reasoned that comparing the two lists could give insight into intestine-specific regulators. Indeed the in3.4 module harbors a number of predicted binding sites not present in the in3.3 region. Overlaying data from truncations and site-directed mutagenesis converged on a set of recognition sites that are within the 40 bp region (henceforth in3.4-CR) required for strong reporter expression (Chapter 3, Figure 3.7B). Most notably, there is a consensus GATA factor binding site within in3.4-CR and no such motif in the in3.3 module. GATA 4, 5, 6 have single zebrafish orthologs and are expressed in the intestinal epithelium [100] at 6dpf. Also of note was a region overlapping the Sub4 mutation that harbored a predicted binding site for nuclear receptors such as Hepatic nuclear factor 4 alpha (Hnf4α), Liver receptor homolog 1 (LRH-1), and Farnesoid Receptor X (FXR). These factors therefore represent potential novel candidate regulators of intestinal expression of *angptl4*.

4.3.2 Substitution of the GATA factor binding site

I next mutated the consensus GATA binding site located within in3.4-CR and assayed this construct for competency to drive intestinal expression in the zebrafish. I found that substitution of only 3 bases within the entire in3.4 module (Figure 4.1A,B) strongly attenuated intestinal reporter expression (Figure 4.1C,D). Interestingly, this GATA motif, and an adjacent region harboring predicted binding sites for other candidate

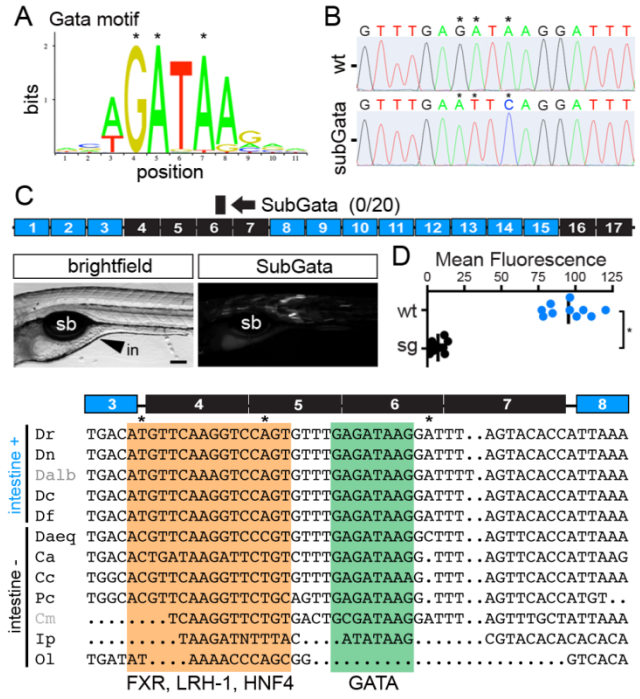


Figure 4.1: Mutation of a predicted GATA factor-binding site abolishes intestinal expression
(A) Logo of the consensus transcription factor binding site for GATA family members from the JASPAR database. Asterisks mark sites of targeted mutagenesis. (B) Sequencing traces showing site-directed mutagenesis substituting 3 base pairs within the 316 bp in3.4 regulatory module comprising the consensus GATA binding site (SubGATA). (C) Representative image of a 7dpf zebrafish injected with the SubGATA mutant construct. Very weak to no intestinal expression was observed in all 20 animals imaged despite moderate muscle expression (likely conferred by the *cfos* minimal promoter). (D) Relative mean intestinal fluorescence within the intestine was quantified in mosaic animals and plotted per injected fish. Circles represent mean fluorescence averaged for three mosaic patches within one fish. Statistical significance was tested using an un-paired Student's T-test (P<.0001). (E) Multiple alignment of intestine positive and intestine negative sequences orthologous to zebrafish in3.4-CR. Alignments of the GATA or Farnesoid Receptor X (FXR), Hepatic nuclear factor-4 (HNF4), Liver receptor homolog-1 (LRH-1) binding sites are highlighted in green or orange, respectively. Asterisks denote bases differentially conserved between intestine positive and intestine negative sequences. *Danio rerio* (*Dr*, zebrafish), *Danio nigrofasciatus* (*Dn*), *Danio albolineatus* (*Dalb*), *Danio choprae* (*Dc*), *Danio feegradei* (*Df*), *Devario aequipinnatus* (*Daeq*, giant danio), *Carassius auratus* (*Ca*, goldfish), *Cyprinus carpio* (*Cc*, carp), *Puntius conchonius* (*Pc*, rosy barb), *Chromobotia macracanthus* (*Cm*, clown loach), *Ictalurus punctatus* (*Ip*, channel catfish), *Oryzias latipes* (*Ol*, medaka). Note that *Dalb* and *Cm* have not been tested (grey).

factors (Hnf4α, LRH-1, FXR) are well-conserved even in Ostariophysi species whose orthologous in3.4 region does not drive intestinal reporter expression (Figure 4.1D). Single-nucleotide differences between intestine-positive and intestine-negative

sequences from other species in this region do not overlap with positions that would be predicted to dramatically alter TFBSs. To be clear, this data does not suggest that these predicted transcription factors, such as GATA factors, are not required for intestinal expression of *angptl4*. However, it highlights the potential distraction of relying on *in silico* predictions to discover factors that regulate a gene through a given *cis*-regulatory module, as the same factors are predicted to bind in sequences that are negative in the intestine. Therefore, I chose to pursue a DNA-centered assay to discover candidate transcription factors that function through in3.4 to regulate intestinal expression of *angptl4* in the zebrafish as a counterpart to protein-centered methodologies.

4.3.3 Generation of an *in vitro* binding assay

I first set out to develop a biochemical binding assay to monitor DNA-protein complex assembly with the goal to partially purify transcription factors functioning through the in3.4-CR regulatory region using DNA affinity chromatography. It was observed that expression of the fluorescent reporter was maintained in adult zebrafish harboring a stable insertion of the in3.2:TdT or in3.4:GFP transgene (Figure 4.2B and data not shown). I therefore inferred that transcription factors should be present and active in these cells sufficient to bind and activate the in3.4-CR element. Compared to the small zebrafish larvae used in the previous chapter, adult intestinal cells present a relatively abundant source material for nuclear extract preparation. Nuclear extracts were prepared from primary intestinal epithelial cells harvested from adult zebrafish intestines (see Methods) and used in non-radioactive electrophoretic mobility shift assays performed with a 50 base pair biotinylated probe (in3.4-CR, Figure 4.2A). I observed at least two specific shifts (termed upper and lower) that were efficiently

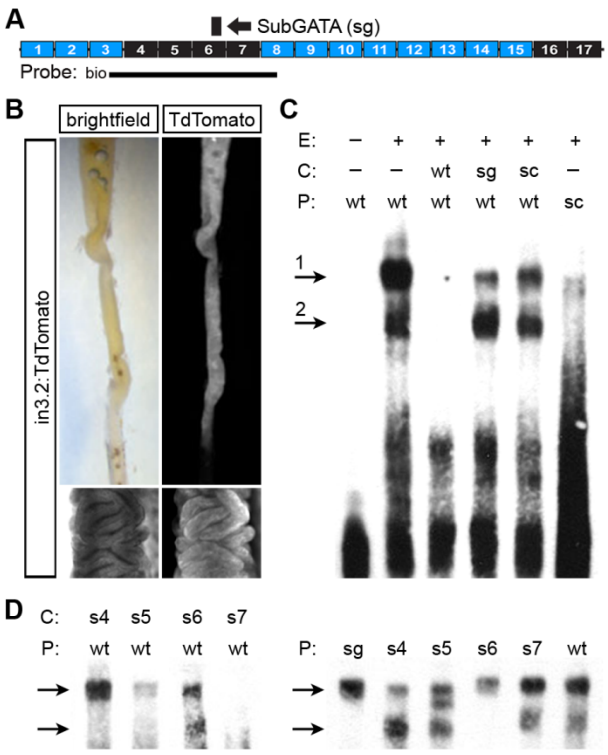


Figure 4.2: Factors in nuclear extracts bind the in3.4-CR regulatory region
(A) Schematic of substitution mutations assayed in Chapter 3 (Figure 3.6). The region used to design the 50 bp biotinylated in3.4-CR probe for the electrophorectic mobility shift assays (EMSA) is shown as a line below the schematic. Mutant probes cover the same region with localization of substitutions indicated. (B) Adult zebrafish maintain expression of the reporter TdTomato (TdT) driven by the in3.2 regulatory module in the intestinal epithelium suggesting transcription factors are present and active. (C) Nuclear extracts from zebrafish adult intestinal epithelial cells harbor factors that bind a 50 bp in3.4-CR double-stranded probe. Arrows mark at least two (upper (1) and lower (2)) complexes. Binding is efficiently competed by 50x unlabeled wild type (wt) DNA, but not with SubGATA (sg) or scrambled (sc) unlabeled DNA. Scrambled probe was generated by scrambling the wild type sequence and does not support complex assembly. Note that similar results were observed with 25x and 100x competitor (not shown). E = extract, C = 50x competitor, P = probe. (D) Mutant DNAs (Sub4, Sub5, Sub6, but not Sub7) are deficient at assembling and competing complexes.

competed away by wild type unlabeled competitor, but not with unlabeled competitor in which the wild type sequence was randomly scrambled (Figure 4.2C). Further, the upper and lower shifts were not apparent when using a biotinylated scrambled probe (Figure 4.2C). These data suggest that the in3.4-CR region can assemble multiple and specific DNA-protein complexes using nuclear extracts from intestinal epithelial cells.

I next assayed the competency of the various substitution mutations to assemble complexes using this electrophoretic mobility shift assay. Strikingly, unlabeled competitor oligos harboring the overlapping subGATA or Sub6 mutation did not fully compete either the upper or lower shift (Figure 4.2C). Interestingly, the upper band still assembled on the subGATA or Sub6 mutant probe but the lower band did not (Figure 4.2D). Furthermore, the Sub4 mutation could efficiently compete away the lower band but not the upper band. Consistent with this observation, the Sub4 mutant could assemble the complex revealed in the lower band, but was moderately deficient in assembling the upper complex. Sub5 appeared moderately deficient in assembling the upper complex, whereas Sub7 was similar to wild type in the ability to assemble and compete for both complexes. This data suggests that the lower shift requires the binding sites affected by the subGATA/Sub6 mutations and the upper shift likely requires the binding sites affected by the Sub4 mutations. Most importantly, the DNA motifs located within the in3.4-CR region are required for both *in vivo* reporter activity and *in vitro* protein-DNA complex assembly.

4.3.4 DNA-affinity chromatography and mass spectrometry to identify transcription factors

Results from the EMSAs suggested that factors within IEC nuclear extracts can specifically assemble on in3.4-CR double-stranded oligos, but EMSA is not well-suited for the direct identification of binding proteins. I attempted supershift experiments using antibodies targeting zebrafish (Dr) GATA binding factors 4, 5, 6 and mouse Rel A of the NFkB complex and these gave inconclusive negative results (data not shown). Therefore, I next sought to establish a DNA-affinity pull-down strategy to discover which protein factors bind the in3.4-CR region. Based on the observations that the subGATA mutation strikingly altered *in vivo* reporter activity and *in vitro* shift activity, I reasoned

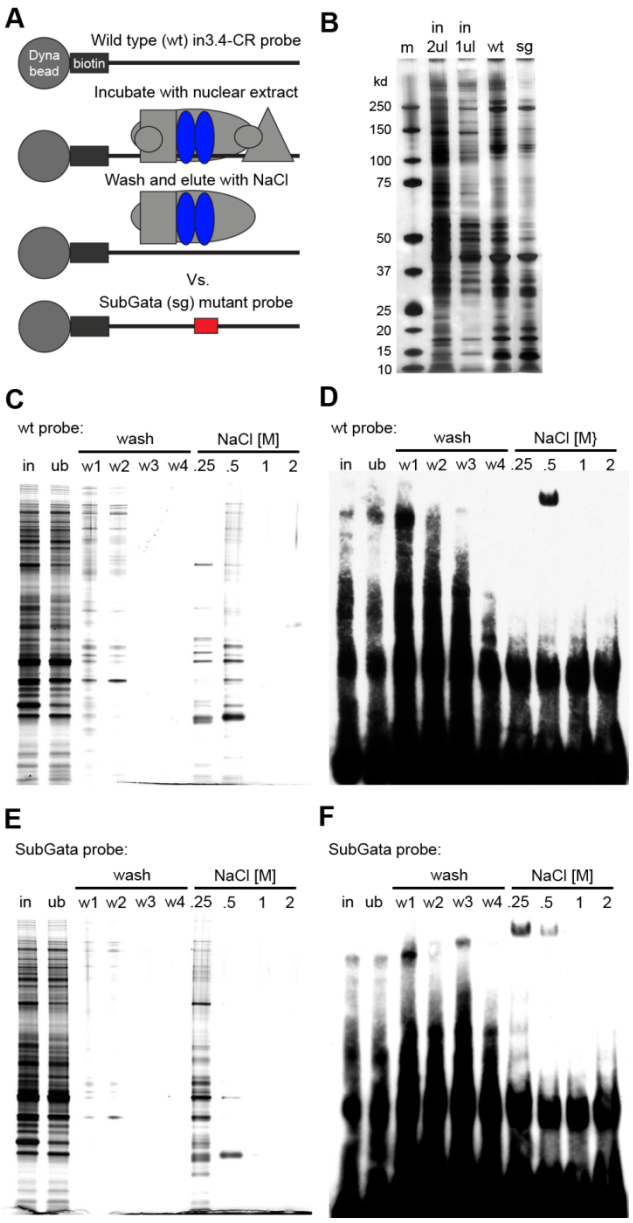


Figure 4.3: DNA affinity pull-down using wild type and subGATA probes
(A) Cartoon schematic showing the pull-down strategy. Biotinylated 50 bp wild type in3.4-CR or SubGATA mutant probe is conjugated to streptavidin-coated magnetic beads and incubated with nuclear extract from zebrafish adult intestinal epithelial cells. Protein-DNA complexes are washed and eluted with various concentrations of NaCl or 2% SDS. (B) Silver stained SDS-polyacrylamide gel showing denatured protein from input extract (in, 1µl and 2µl to show strong and weak bands), bound protein from wild type pull-down (wt) and from SubGATA pull-down (sg). Protein was eluted with 2% SDS at 85°C for 10 minutes to recover all bound protein. m = marker (kb). (C-D) The experiment was repeated this time eluting with increasing molar concentration of NaCl. (C) Each fraction (in = input nuclear extract, ub = protein unbound after 4 hour incubation, w = wash) was separated using SDS-PAGE and silver stained or (D) dialyzed and assayed for binding activity to the wild type probe through a non-radioactive electromobility shift assay (EMSA). Note the strong shift activity observed in the 500 mM NaCl fraction. (E-F) The experiment was repeated with the SubGATA probe. Note that the shift activity apparently elutes at a lower NaCl concentration though this particular observation has not been repeated.

that this probe could be useful in comparison to wildtype in3.4-CR. In the first iteration, a biotinylated 50 base pair wild type or subGATA probe was linked to magnetic beads via streptavidin and incubated with nuclear extracts from primary zebrafish intestinal epithelial cells for 4 hours at 4°C. Bead-DNA-protein complexes were washed multiple times with binding buffer containing non-specific poly (dI-dC) and the remaining bound protein was eluted with 2% SDS at 85°C for 10 minutes, separated using SDS-PAGE, and silver-stained (Figure 4.3B). This revealed many protein bands most of which were present in both the wild type and subGATA lanes. In order to further distinguish potential differences between the wild type probe and the subGATA probes, I next washed Bead-DNA-protein complexes 4 times with binding buffer containing non-specific poly dI/dC and eluted with increasing concentrations of NaCl. Input extract, unbound extract, washes and salt elutions were first dialyzed, then (i) separated using SDS-PAGE and silver-stained or (ii) assayed for competency to shift wild type double-stranded in3.4-CR probe using EMSA. After the 3rd wash there was no detectable protein in the elution and increasing the salt concentration stepwise dissociated bound protein (Figure 4.3C). Interestingly, there was a strong shift in only one lane (0.5M NaCl) from the wild type pull-down elutions (Figure 4.3D), which correlated with a number of strong bands present in the corresponding 0.5M NaCl elution lane of the silver-stained gel. The analogous experiment with SubGATA mutant probe had a similar shift though this activity was eluted with a weaker salt elution (Figure 4.3E,F). Note that it is unclear if the shift activity in these experiments is the same activity observed in the previous EMSA (Figure 4.2).

To ascertain the specificity of the EMSA activity and protein bands eluted in the wild type and SubGATA experiments, I repeated the pull-downs using a scrambled

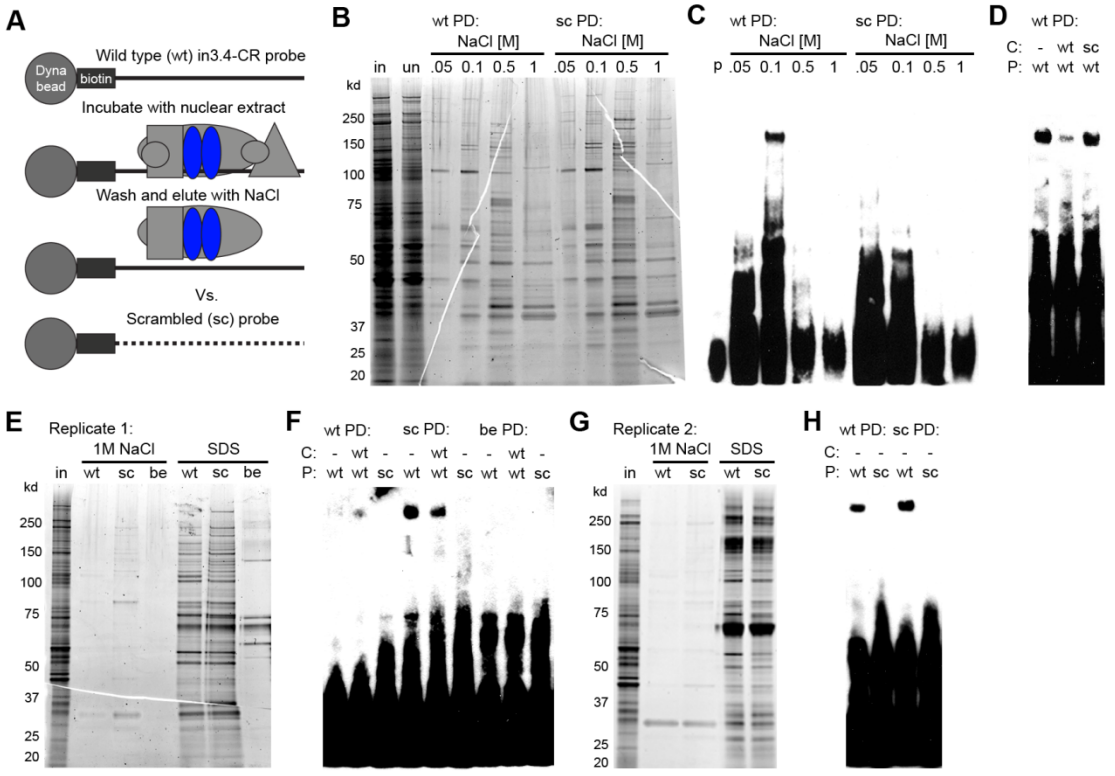
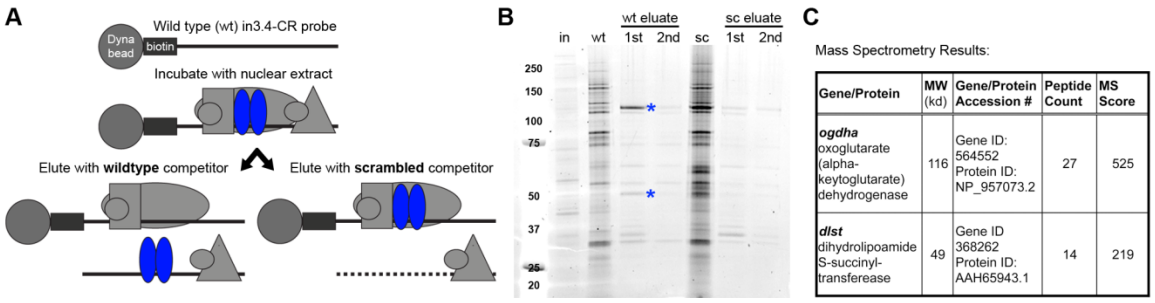


Figure 4.4: DNA affinity pull-down using wild type and scrambled probes
(A) Cartoon schematic showing the pull-down strategy. Biotinylated 50 bp wild type in3.4-CR or scrambled probe is conjugated to streptavidin-coated magnetic beads and incubated with nuclear extract from zebrafish adult intestinal epithelial cells. Protein-DNA complexes are washed and eluted with various concentrations of NaCl or 2% SDS. (B) Sypro-ruby stained SDS-polyacrylamide gel showing denatured protein after eluting with increasing molar concentration of NaCl from wild type (wt PD) or scrambled (sc PD) DNA-bead complexes (in = input nuclear extract, ub = protein unbound after 4hour incubation with wt probe). (C) Each salt fraction was dialyzed and assayed for binding activity to the in3.4-CR wild type probe using the non-radioactive electromobility shift assay (EMSA). Note the strong shift activity observed in the 100 mM NaCl fraction from the wild type pull-down. (D) The shift activity from the wt pull-down was efficiently competed by wild type unlabeled competitor (50x) but not scrambled unlabeled competitor (50x). (E-H) The experiment was repeated twice (Replicate 1 - E,F; Replicate 2 - G,H) this time adding BSA and scrambled competitor into binding reactions and wash buffer, eluting once with 50 μl of 1 M NaCl at room temperature for 3 min, and a second time with 50 μl of buffer containing 2% SDS at 85°C for 10min. Beads only (be) bind few proteins (E) and elicit no shift (F). However, in these replicates a shift was observed using eluate from the scrambled pull-downs (F,G). The shift activity is not seen in an EMSA using scrambled probe with eluates from wildtype and scrambled pull-downs (H). Note that the shift was also present in the wild type pull-down though to a lesser extent (F, H).

probe (Figure 4.4A). The resultant SDS-PAGE separation and SyproRuby stained gel from wild type or scrambled revealed a strikingly similar pattern of protein bands upon step-wise elution with increasing NaCl concentrations (Figure 4.4B). Encouragingly, I observed a single strong shift activity in the 0.1M NaCl elution in the wild type pull-down,

but no shift activity in any fraction from the scrambled probe (Figure 4.4D). Also, this shift activity was competed away by wild type unlabeled competitor, but not with scrambled unlabeled competitor (Figure 4.4D). Despite this specific shift activity on the wild type probe, there appeared to be no overt differences in the SyproRuby stained gel from the wild type pull-down compared to the scrambled probe pull-down that correlated with this shift activity (Figure 4.4B). It is likely that there are low abundant proteins differentially present in the wild type and scrambled pull-downs. These initial results were very promising, but suggested that many proteins were binding to the probe and/or beads in a non-specific manner. This prompted the need to de-noise the pull-downs. In one attempt I included excess scrambled oligo into the reactions and a short (3 minute) wash step with high salt. In this experiment I observed a similar set of proteins in wild type and scrambled pull-downs (Figure 4.4E,H), however the shift activity was now weak in the wildtype pulldown and present in the scrambled pull-down (Figure 4.4F,H). This could be explained by experimental error, but I replicated the same results in an independent experiment (Figure 4.4G,H). Notably, eluate from bead only control pull-downs had few proteins and did not elicit a shift (Figure 4.4E,F). Curiously, eluate from wild type and scrambled pull-downs shifted only the wild type probe and did not shift the scrambled probe (Figure 4.4F,H). This data suggested that further optimization or alternative strategies should be pursued to enrich for wild type-specific DNA-binding proteins in order to differentially distinguish protein bands using gel electrophoresis and EMSA activity.

In an alternative strategy, I utilized the observation from previous EMSA assays that incubation with unlabeled competitor DNA can be used to specifically compete DNA binding factors (Figure 4.2). In this iteration, the strategy was to attach the wild type probe to magnetic beads, incubate with nuclear extract, wash with binding buffer, and elute with either wild type or scrambled unlabeled competitor (Figure 4.5A). DNA-protein



Gene/Protein	MW (kd)	Gene/Protein Accession #	Peptide Count	MS Score
ogdha oxoglutarate (alpha-keytoglutarate) dehydrogenase	116	Gene ID: 564552 Protein ID: NP_957073.2	27	525
dlst dihydrolipoamide S-succinyl-transferease	49	Gene ID 368262 Protein ID: AAH65943.1	14	219

Figure 4.5: DNA affinity pull-down using wild type and scrambled competitors
(A) Cartoon schematic showing the pull-down strategy. Biotinylated 50 bp wild type in3.4-CR probe is conjugated to streptavidin-coated magnetic beads and incubated with nuclear extract from zebrafish adult intestinal epithelial cells. Protein-DNA complexes are washed and eluted with 50x wild type or scrambled competitor. (B) SDS-PAGE SyproRuby stained gel of input extract (in), protein unbound to the wild type DNA-bead complex after incubation (ub), protein still bound to the wild type DNA-bead complex after a two elutions with 50x wild type (wt), the 1st elution with wt competitor (1st), the 2nd elution with wt competitor (2nd), protein still bound to the wild type DNA-bead complex after a two elutions with 50x scrambled competitor (sc), the 1st elution with sc competitor (1st), the 2nd elution with sc competitor (2nd). (C) Mass spectrometry identification of the protein bands marked with a blue asterisks in (B).

complexes remaining on the beads after elution are then denatured and eluted with 2% SDS at 85°C for 10min. Here, one would expect factors specific for the wild type sequence to be present in fractions representing the wild type competitor elution rather than remaining bound to the bead-DNA complex. Indeed, this resulted in at least two bands that were enriched when eluted with wild type competitor in comparison to scrambled competitor (Figure 4.5B). Furthermore, the same proteins were still present on the wild type bead-DNA complex after the elution with scrambled competitor, but not with wild type competitor (Figure 4.5B). We successfully identified two of the bands by mass spectrometry as oxoglutarate dehydrogenase (*Ogdha*) and dihydrolipoamide S-succinyl-transferase (*Dlst*), both of which are components of the oxoglutarate dehydrogenase complex. I had previously identified one of these components (*Ogdha*) in another iteration of this pull-down in which I performed the incubations in the presence of wild type or scrambled competitor (data not shown). Unfortunately, neither of these

proteins has any known DNA-binding properties and the complex has well characterized enzymatic functions in the citric acid cycle [277].

It is possible that transcription factors are enriched in the in3.4-CR pulldowns compared to scrambled, but are not apparent in the SDS-PAGE gel. It has been reported that low abundance transcription factors enriched using DNA affinity pull-down assays can be identified even in complex mixtures [275,276]. I next returned to the wildtype vs. scrambled pulldown strategy (Figure 4.4), but this time I included stringent washing steps using low salt (50 mM NaCl) and weak detergent (0.01% Triton X-100) in an attempt to remove non-specific proteins from the wild type and scrambled pull-downs. In an effort to ensure sufficient protein for mass spectrometry, I did not perform SDS-PAGE and used all of the wild-type in3.4 pull-down and scrambled pull-down for in solution digestion and peptide identification by LC MALDI-TOF/TOF. We were able to identify only a small number of proteins with confidence (Table 4.1, 4.2). Although there were a number of differences in proteins identified between wild type and scrambled pull-downs, there was no striking sequence-specific transcription factor discovered. Furthermore, the amount of total peptides identified by mass spectrometry was very low suggesting that the washing conditions were perhaps too stringent or the method of peptide extraction from the beads was not optimal.

Altogether, the above data provides the foundation for multiple methodologies aimed towards the identification of transcription factors that function through the *angptl4* in3.4 regulatory module. Strategies for improvement are discussed extensively below.

Wildtype Pull-down:

IPI Accession #	Abreviation	Name	Genbank Accession #	Peptide Count (95%)
IPI:IPI00771914.2	LOC100332815	Hypothetical protein LOC100332815	XP_002666966	8
IPI:IPI00654475.1	hnrnpu	Heterogeneous nuclear ribonucleoprotein U	NP_001028767.2	7
IPI:IPI00490006.2	bactin2	Actin, cytoplasmic 2	NP_853632.2	7
IPI:IPI00496722.7	rrbp1a	Ribosome binding protein 1 homolog a	NP_001116531.1	5
IPI:IPI00483673.3	LOC560910	Hypothetical protein LOC560910 isoform 6	XP_708065	5
IPI:IPI00932198.1	LOC565341	Hypothetical protein LOC565341	XP_693710	5
IPI:IPI00495050.6	wu:fc51c09	Heterogeneous nuclear ribonucleoprotein U	P_694691	3
IPI:IPI00570342.2	zgc:110380	Uncharacterized protein	CT663081	2
IPI:IPI00507036.2	sfpq	Splicing factor proline/glutamine-rich	NP_998443	2
IPI:IPI00508405.4	acad11	member 11	NP_956472.1	2
IPI:IPI00860024.2	top1l	Topoisomerase (DNA) I, like	NP_001037789.1	2
IPI:IPI00607268.1	zgc:110425	Uncharacterized protein	Zgc:110425	2
IPI:IPI00994952.1	zgc:153867	isoform	NP_998803	1
IPI:IPI00638238.5	myo6a	Myosin 6a	NP_001004111	1
IPI:IPI00617405.4	LOC560949	Uncharacterized protein	AAH91989	1
IPI:IPI00500109.4	jazf1a	Juxtaposed with another zinc finger protein 1	NP_001038420.1	1
IPI:IPI01007476.1	si:ch211-242b18.1	Uncharacterized protein	CA471455	1
IPI:IPI00492110.2	slc25a5	ADP/ATP translocase 2	NP_775354	1
IPI:IPI00837436.2	ccdc12	Coiled-coil domain containing 12	NP_001003589.1	1

Scrambled Pull-down:

IPI Accession #	Gene Symbol	Protein Name	Genbank Accession #	Peptide Count (95%)
IPI:IPI00961172.1	LOC100334868	Hypothetical protein LOC100334868	XP_002666950	6
IPI:IPI00637460.2	mccc2	Methylcrotonoyl-Coenzyme A carboxylase 2	NP_998092.1	5
IPI:IPI00899187.2	wu:fd60h05	Uncharacterized protein	AW018852	3
IPI:IPI00490006.2	bactin2	Actin, cytoplasmic 2	NP_853632.2	3
IPI:IPI00505686.2	LOC562307	Hypothetical protein LOC562307 isoform 2	XP_708197.1	3
IPI:IPI00496722.7	rrbp1a	Ribosome binding protein 1 homolog a	NP_001116531.1	5
IPI:IPI00932245.1	mccc1	(LOC559969)	NP_001071208.1	2
IPI:IPI00897805.2	zgc:111961	ATP synthase subunit beta	NP_001019600	2
IPI:IPI00490232.4	ckmt1	Creatin kinase U-type, mitochondrial	NP_942096	2
IPI:IPI00483853.1	glud1b	Glutamate dehydrogenase 1b	NP_955839.2	2
IPI:IPI00491975.2	atp5a1	ATP synthase subunit alpha, mitochondrial	NP_001070823.1	2
IPI:IPI00570342.2	zgc:110380	Uncharacterized protein	CT663081	2
IPI:IPI00607268.1	zgc:110425	Uncharacterized protein	Zgc:110425	2
IPI:IPI00851936.2	LOC569982	Histone H1-like LOC569982	XP_698497	1
IPI:IPI00512240.1	ef1a	Elongation factor 1-alpha	NP_571338.1	1
IPI:IPI00492110.2	slc25a5	ADP/ATP translocase 2	NP_775354	1
IPI:IPI00953828.2	pkhd-l1, si:ch211-244a23.1	Polycysti kidney and hepatic disease-like 2	ABG74915.2	1
IPI:IPI00499836.5	papss2a	synthase 2a	NP_001071235	1
IPI:IPI00960461.2	LOC100330159	Glutamine-rich protein 2-like	XP_002661338	1
IPI:IPI00833098.3	itpr3 U	Inositol 1,4,5-triphosphate receptor, type 3	NP_001121741	1
IPI:IPI00500109.4	jazf1a	Juxtaposed with another zinc finger protein 1	NP_001038420.1	1
IPI:IPI00489853.1	rps10	Ribosomal protein S10	NP_957440.1	1
IPI:IPI00837436.2	ccdc12	Coiled-coil domain containing 12	NP_001003589.1	1
IPI:IPI01007476.1	si:ch211-242b18.1	Uncharacterized protein	CA471455	1
IPI:IPI00835722.2	zgc:158846	Zgc:158846 protein	NP_001076492.1	1
IPI:IPI00896289.1	smarcal1	SWI/SNF-related matrix-associated actin-dependent regulator of chromatin subfamily A-like protein 1	NP_001120938.1	1

Table 4.1: Mass spectrometry results from wild type and scrambled pull-downs using IPI
Proteins identified from wild type (top) and scrambled (bottom) pull-downs using LC-MALDI-TOF/TOF are shown. Peptides were searched against the International protein Index (IPI) database for zebrafish. Highlighted in gray are the proteins common to both wild type and scrambled pull-downs.

Wildtype Pull-down:

UniProt Accession #	Abbreviation	Name	Species	Peptide Count (95%)
P02769	ALB	Serum albumin	*Bos taurus*	11
Q7ZVF9	actbb	Actin, cytoplasmic 2	*Danio rerio*	4
P00761	Prss-1	Trypsin	*Sus scrofa*	4
P06350	His1	Histone H1	*Oncorhynchus mykiss*	2
P22629	Strav	Streptavadin	*Streptomyces avidinii*	2
Q9MYN8	TTR	Transthyretin	*Erinaceus europaeus*	1
P04264	KRT1	Keratin, type II cytoskeletal 1	*Homo sapiens*	1
Q99K48	Nono	Non-POU domain-containing octamer-binding protein	*Mus musculus*	1
Q6IFW6	Krt10	Keratin, type I cytoskeletal 10	*Rattus norvegicus*	0
Q5F928	uvrC	UvrABC system protein C	*Neisseria gonorrhoeae*	0
Q4ZMG5	rsmA	Ribosomal RNA small subunit methyltransferase A	*Pseudomonas syringae*	0
Q0BWA9	mnmG	tRNA uridine 5-carboxymethylaminomethyl modification enzyme	*Hyphomonas neptunium*	0
Q99RB9	fbp	Fructose-1,6-bisphosphatase class 3	*Staphylococcus aureus*	0
Q251S4	pth	Peptidyl-tRNA hydrolase	*Desulfitobacterium hafniense*	0
Q9HZA3	leuC	3-isopropylmalate dehydratase large subunit	*Pseudomonas aeruginosa*	0
Q796Q6	yisV	Uncharacterized HTH-type transcriptional regulator yisV	*BACSU*	0
C4QWJ4	MDM10	Mitochondrial distribution and morphology protein 10	*Pichia pastoris*	0
Q80XL6	Acad11	Acyl-CoA dehydrogenase family member 11	*Mus musculus*	0
P10590	Ppst	Putative packaging signal terminase	*Vibrio phage*	0
Q86WI3	NLRC5	NOD-like receptor C5	*Homo sapiens*	0
P21259	Prfa	Pol-RFamide neuropeptides OS	*Polyorchis penicillatus*	0
P17867	cisA	Putative DNA recombinase	*Bacillus subtilis*	0
Q2NE10	ilvD	Dihydroxy-acid dehydratase	*Methanosphaera stadtmanae*	0
Q69YN4	KIAA1429	Protein virilizer homolog	*Homo sapiens*	0

Scrambled Pull-down:

UniProt Accession #	Abbreviation	Name	Species	Peptide Count (95%)
P22629	Strav	Streptavidin	*Streptomyces avidinii*	1
Q9Y707	ACT	Actin-2	*Suillus bovinus*	1
A1CRE5	tif32	Eukaryotic translation initiation factor 3 subunit A	*Aspergillus clavatus*	0
C4QWJ4	MDM10	Mitochondrial distribution and morphology protein 10	*Pichia pastoris*	0
Q83BR3	nuoI	NADH-quinone oxidoreductase subunit I	*Coxiella burnetii*	0
Q4FZU2	Krt6a	Keratin, type II cytoskeletal 6A	*Rattus norvegicus*	0
Q6IFW6	Krt10	Keratin, type I cytoskeletal 10	*Rattus norvegicus*	0

Table 4.2: Mass spectrometry results from wild type and scrambled pull-downs using UniProt
Proteins identified from wild type (top) and scrambled (bottom) pulldowns using LC-MALDI-TOF/TOF are shown. Peptides were searched against the UniProt database which include all species (including bacteria). Highlighted in gray are the proteins common to both wild type and scrambled pull-downs.

4.4 Discussion

4.4.1 Potential transcription factors regulating intestinal transcription of *angptl4*

Sequence-specific transcription factors select genes for activation or repression through direct interaction with DNA regulatory sequences. It remains a difficult task to identify transcription factors regulating a gene of interest given a known *cis*-regulatory module. I have defined a minimal region within the in3.4 CRM that harbors regulatory activity in the intestine that is conserved within the *Danio* lineage (Chapter 3). Predicted transcription factor binding sites within this region intimates potential roles for these factors in regulation of *angptl4* tissue-specific transcription and/or microbial suppression. Because sequence-specific transcription factors typically recognize 6-12 bp motifs [222], it is reasonable to assume that multiple factors cooperate to combinatorially regulate intestinal expression through this CRM. The Hnf4 family of fatty acid-regulated nuclear receptors has evolutionarily conserved roles in lipid metabolism [223], and Hnf4α is expressed in the intestinal epithelium of zebrafish [99] and mouse [143]. Similarly, GATA factors 4, 5, and 6 are all expressed in the zebrafish [99,100] and mouse [97,98] intestinal epithelium and have proposed roles in regulating epithelial cell differentiation. Notably, *C. elegans* GATA family member *elt-2* has been implicated in mediating intestinal epithelial cell immune responses [102] suggesting that GATA factors could mediate tissue-specific as well as microbial regulatory inputs at *angptl4*. Indeed, mutation of the predicted GATA binding site within in3.4-CR abrogated intestinal expression (Figure 4.1C). PPAR family members have been identified as key regulators of mammalian *Angptl4* expression in adipocytes and hepatocytes through PPAR responsive elements located in the 5' portion of human *ANGPTL4* intron 3 [177,180] and zebrafish PPARγ [226] and PPARδ [227] homologs are expressed in the larval intestine. The zebrafish *angptl4* locus contains multiple predicted PPRE sites, including several in

both the 5' and 3' portion of intron 3 [228]. Most notably, a predicted PPRE was detected within the substitution blocks 16/17 in the intestinal enhancer in3.4 (Figure 3.7B). However, the PPREs within zebrafish *angptl4* intron 3 that display the highest sequence homology to the defined human *ANGPTL4* intron 3 PPRE mapped outside of minimal regions for either intestinal or islet expression within the 5' liver module (data not shown). The location of these PPREs in the 5' region of zebrafish *angptl4* intron 3, combined with the fact that the PPREs discovered in human *ANGPTL4* are also located in the 5' portion of intron 3, suggests that the predicted PPREs within the 3' islet and intestine CRMs of zebrafish *angptl4* could represent novel elements for which functional equivalents have not been identified in mammals.

Although these predicted factors represent candidates for controlling intestine-specific regulation of *angptl4*, databases of predicted TFBSs are incomplete and commonly produce both false-positive and false-negative predictions. Moreover, critical regions identified by SDM might reflect sequences that alter nucleosome positioning or histone modification patterns rather than binding sites for sequence-specific transcription factors. However, the correlation of *in vivo* and *in vitro* results argues against this possibility. I anticipate that unbiased methods for transcription factor discovery will provide the most rigorous approach to an improved understanding of the *angptl4* *cis*/*trans* program.

4.4.2 Optimization of methods for the unbiased discovery of transcription factors

The zebrafish has extraordinary potential as a vertebrate model organism to discover transcriptional programs regulating intestinal physiology and pathophysiology, however methods for identifying and studying transcription factors in the zebrafish intestine are not well established. In order for the zebrafish model to be maximally useful for medium to high-throughput discovery and characterization of *cis*/*trans* regulation of

gene expression, then unbiased methods for transcription factor discovery should be pursued. In this work, we adapted classical biochemical assays, namely EMSA and DNA affinity chromatography, for use with extracts from zebrafish intestinal epithelial cells. These assays have been instrumental in other systems for characterizing and isolating the DNA binding activity of factors present in a given cell type [270-272,275,276]. Here I show that factors present in adult zebrafish intestinal epithelial cells can assemble specific complexes on the in3.4-CR sequence. Intriguingly, the putative transcription factor binding sites located within the in3.4-CR region that were originally defined through functional truncation assays, site-directed mutagenesis, and evolutionary conservation *in vivo* are also required for *in vitro* protein-DNA complex assembly. These findings (i) strengthen the support for the critical role of this intronic region in regulating *angptl4* intestinal transcription, and (ii) foster confidence in the general utility of these two complementary approaches.

I used binding conditions established in the EMSA experiments to attempt partial purification of transcription factors using DNA affinity chromatography and employed mass spectrometry to identify peptides present in the eluate. This effort led to the positive identification of many factors including two components of the oxoglutarate dehydrogenase complex, namely oxoglutarate dehydrogenase (Ogdha) dihydrolipoamide S-succinyl-transferase (Dlst). Further, Ogdha was identified as differentially present in the wild type pull-down using two independent experimental strategies. Multiple copies of Ogdha, Dlst, and dihydrolipoamide dehydrogenase (Dld) proteins form a mitochondrial complex that catalyzes the conversion of 2-oxoglutarate (alpha-ketoglutarate) to succinyl-CoA and carbon dioxide during the Krebs cycle. I was unable to identify any published reference of DNA-binding activities of any of these proteins in the literature. It is indeed interesting that this complex readily elutes upon addition of unlabeled wild type competitor, but not with unlabeled scrambled competitor.

It is plausible that we have uncovered a novel function of an ancient metabolic enzyme normally localized to the mitochondria, however it is more likely that this is a spurious binding event. P-fam and PROSITE databases both predict that Dlst has a biotin/lipoyl attachment domain, which raises the possibility tht the streptavidin/biotin method used for coupling DNA to beads promote binding of Dlst and other complex members. These and other concerns highlight the need to further optimize pull-down conditions. In general, close collaboration or consultation with experts in transcription factor purification using DNA chromatography is strongly recommended. Other suggestions for improvements, considerations, and alternative approaches are highlighted below:

i) One of the major limitations in these experiments was extract quantity. In hindsight, transcription factors are thought to be in low abundance in cells, and it would therefore be optimistic to expect observable enrichment from a single chromatographic step on a protein-stained gel. It is more likely that hidden amongst the bands, or invisible on the gel, is a low copy number transcription factor that would be differentially present in wild type and mutant pull-downs. Therefore, if extract quantity cannot be increased to allow for tandem enrichment, then a one-step pull-down followed by identification in complex mixtures (such as multidimensional protein identification technology, MudPIT) could be the best way forward. Multiple replicates using multiple mutant or scrambled controls would help filter the abundant non-specific binding proteins.

ii) Due largely to inexperience, we used a commercial kit to generate nuclear extracts. Though the methodology was similar to that described [278,279], it is unclear how well this kit enriched for nuclear proteins. It would be useful to perform a western blot for nuclear proteins such as histones in the nuclear extract versus the cytosolic fraction. I performed a western blot for GATA factors

4,5,6 and this showed enrichment in the nuclear fraction, however I did not control for input protein (data not shown). Furthermore, components of the buffer solutions in the commercial kit are proprietary and this is not conducive to scaling this protocol or detailed understanding of binding affinity requirements.

iii) We appreciated the elegance of using zebrafish intestinal epithelial extracts as source material for nuclear extract preparation. However, it may be wise to utilize intestines from larger teleost fish, or even mammals, in order to increase the amount of available extract. If similar binding activities were present in these extracts then extract quantity would no longer be limited. It remains an interesting and untested question how the zebrafish in3.4 enhancer would function in another teleost or mammalian species.

iv) The experiments presented in this chapter suggest that there is potential in these methodologies for transcription factor discovery, however the source of nuclear extract is accompanied by several potential caveats. Intestinal epithelial cells synthesize numerous proteases and nucleases [280], and it remains unclear how these enzymes affect extract quality and DNA probe integrity. Protease inhibitors are added to all reactions and proteins in the extracts do not appear by SDS-PAGE separation and staining to be overtly degraded. However, it has become apparent that Dnases are active in sloughed epithelium (see Chapter 5) and may impact interpretation of EMSA and DNA pull-down results.

v) EMSA and pull-down conditions can and should be varied in order to understand the requirements for binding. I tested multiple binding conditions such as addition of glycerol, BSA, salmon sperm, detergent, and bivalent cations MgCl2 and ZnCl2, which either had negligible or deleterious effects on EMSA activities. However, these conditions were never exhausted or replicated and each variable

represents potential improvements in the DNA pull-down, especially in the washing steps.

vi) Extracts were pre-incubated with streptavidin beads to deplete the extracts of non-specific bead or streptavidin binding proteins. It may be advantages to also deplete the extract of biotin binding proteins. It may also be useful to move away from biotin as a method to visualize shifted probes in EMSA and to attach oligos to beads in pull-down strategies. Dlst and likely other enzymes identified may use biotin as a co-factor.

vii) Alternative unbiased approaches should also be considered. Novel high-throughput yeast one-hybrid systems have recently been developed for *C. elegans* [281], Drosophila [282], and human [283]. Such an investment into library development for the zebrafish transcription factor repertoire would greatly enhance the utility of this organism for high-throughput characterization of transcriptional regulatory programs involved in development and disease.

viii) Large scale efforts at defining the binding specificity of TFs should at some point saturate and refine PWMs [67,284]. Though DNA-binding domains are often highly conserved, it will be interesting to observe how useful these data sets defined in mammals will be to zebrafish. One would hope the computational predictive capacity of these databases could become even more reliable, and screening of candidate factors may be the most efficient first step. Standard zebrafish methods for controlling transcription factor activity such as morpholino knockdown and RNA injections have limited utility in the 4-6dpf zebrafish because of diluted effects at this time point and the requirement of many transcription factors during early development. Therefore, reagents such as Vivo-morpholinos [285] or methods in which knockdown or overexpression can be encoded in DNA, including RNAi and inducible expression vectors, could bypass

the obstacle that many transcription factors are required during embryogenesis. These methods would also extend knockdown potential to later points in development. Furthermore, the continued development of targeted knockout approaches using Zinc-finger nucleases [286,287] or TALENS [288,289] may alleviate the problems associated with inadequate knock-down efficiencies.

4.5 Materials and Methods

4.5.1 Zebrafish husbandry

All experiments using zebrafish were performed in wild type TL strains according to established protocols approved by the Animal Studies Committee at the University of North Carolina at Chapel Hill.

4.5.2 Site-directed mutagenesis

To create site-directed substitution of the GATA binding site, two 50 bp complementary primers (Table 4.3) containing a 3 bp mismatch to wild type in3.4-CR were used in a circular PCR of the reporter vector (*Tg(in3.4-Mmu.Fos:GFP)*) followed by DpnI treatment to digest methylated parent plasmid. Nucleotides selected for exchange aimed to maximally disrupt the GATA motif while simultaneously incorporating an EcoRI restriction site in order to screen for mutant bacterial colonies. All plasmids were verified by Sanger dideoxy terminator sequencing.

4.5.3 Injections, imaging, and reporter quantification

Co-injections of Tol2 plasmid and transposase mRNA were performed as described [196]. Generally, 100-200 zebrafish embryos were injected at the 1-2 cell stage with approximately 69 pg of plasmid DNA at a DNA:transposase ratio of 1:2. Injections of the wild type and subGATA mutant constructs were performed with two

sequence-verified plasmids in two independent experiments. Mosaic expression patterns were quantified as follows: at least 200 fish were visually observed and at least 10 fish injected with wild type or subGATA reporter constructs were imaged at the same magnification and exposure time and densitometric measures were quantified in 8-bit grey scale images using ImageJ software [244]. Three mosaic patches within a given tissue of an imaged fish were quantified for mean fluorescence intensity and averaged. Statistical significance was tested using unpaired Student's T-test using GraphPad Prism software.

4.5.4 DNA sequence analysis

DNA sequence from 12 fish species was acquired and analyzed as described in Chapter 3 and Camp et al. 2012 [290].

4.5.5 Motif and transcription factor binding site (TFBS) predictions

DNA sequences were queried for predicted transcription factor binding sites deposited in TRANSFAC [237] and JASPAR [238] databases using MATCH Chekmenev [239] and TESS [240] programs using default settings. We also used a discriminative motif MEME [241] search to discover motifs common to islet-positive or intestine-positive intronic regions, using sequences orthologous to in3.4 or sequences orthologous to in3.3, respectively, as negative selectors. To determine if MEME motifs were unique to islet- or intestine-positive regions, we used MAST [242] to query islet-negative (*Ol* in.3) or intestine-negative (*Daeq, Ca, Cc, Pc, Cm, Ip, Ol* in3.4) sequences for islet-positive or intestine-positive MEME motifs, respectively. TOMTOM [243] was used to query MEME hits against TRANSFAC and JASPAR databases.

4.5.6 Nuclear extracts preparation

Intestinal epithelial cells (IECs) were isolated and nuclear protein extracted mostly as described [291]. Briefly, 3 adult fish intestines were dissected, splayed open, and washed thoroughly with cold PBS. The caudal-most region corresponding roughly to segment 6 and segment 7 [292] were removed. Care was taken to work quickly and remove as much non-intestinal tissue as possible. Intestines were then incubated in Dissociation Reagent 1 (30 mM EDTA, 1.5 mM DTT, 0.5x Complete protease inhibitors (Roche), in 1x PBS) for 15 minutes on ice, then transferred to Dissociation Reagent 2 (30 mM EDTA, 0.5x Complete protease inhibitors (Roche), in PBS) at room temperature and manually shaken for up to 10 minutes to dissociate the epithelial layer. Epithelial cells were collected by pouring into a 15 ml conical tube, pelleted at 500 x G for 5 min at 4 degree, and washed one time with cold 1x PBS. Nuclear protein was extracted using the ActivMotif kit according to the manufacturers specifications for 2 x 10^7 cells, except the hypotonic buffer volume was doubled (2ml). Protein concentration was determined using standard Bradford assays (Invitrogen) [293]. Due to the stickiness of intestinal epithelial cells, many cells were lost to the walls of tips, tubes, and pipettes and precaution was taken to reduce pipetting to a minimum. Cell number per intestine was difficult to estimate because epithelial cell layers were not fully dissociated. The amount of nuclear extract per intestine varied, but ranged from approximately 10-50 μg/intestine.

4.5.7 Electrophoretic mobility shift assays (EMSAs)

Non-radioactive EMSA was performed using the LightShift Chemiluminescent EMSA kit (Pierce). Biotin 5' end-labeled oligo was annealed with unlabelled complementary oligo in oligo annealing buffer (10 mM Tris, 1 mM EDTA, 50 mM NaCl, pH 8) using a thermocycler (Eppendorf) program (heat 95° for 5min, decreasing

temperature to 25° at 1°/minute). EMSA binding reactions using 1x EMSA binding buffer (10 mM Tris-HCl, 50 mM KCl, 1 mM DTT, pH 7.5), 100 ng/μl Poly (dI-dC), IEC nuclear extract (0.1-0.5 μg/μl), were pre-incubated for 10 minutes at room temperature. Labeled probe was added (2 fmol/μl) and the reactions were incubated at room temperature for 20 minutes prior to loading 5μl into a 6% polyacrylamide gel DNA-retardation gel (Invitrogen EC63655BOX). In competition experiments unlabelled oligos were annealed in oligo annealing buffer and pre-incubated (50 fmol/μl, 100 fmol/μl or 200 fmol/μl) with extracts for 10 min prior to addition of labeled probe. Reactions were electrophoresed for 80 min at 100 volts and transferred to positively charged nylon membranes (GE Healthcare) for 45 minutes at 380 Amps. Visualization of band shifts were carried out exactly as described according to manufacturer specifications (Pierce). Sequences of wild type, mutant, and scrambled oligos used in EMSA are listed in Table 4.3.

4.5.8 DNA affinity chromatography assays

Assay conditions for the DNA affinity pull-down experiments were similar to those established in the EMSA experiments but varied slightly amongst the different experiments presented in this thesis. In all experiments DNA probes and competitors were annealed as described above. 10 μg (approximately 300 pmol) of double stranded wild type, mutant, or scrambled DNA probe was coupled to 1 mg of Dynabead M-280 streptavidin coated beads (Invitrogen) in 1x B&W buffer (5 mM Tris-HCl (pH 7.5), 0.5 mM EDTA, 1 M NaCl) per the manufacturer's specifications. Zebrafish IEC nuclear extracts were pre-cleared with 0.25 mg washed streptavidin-coated beads for 10 minutes with gentle rotation at 4°C and centrifuged at 20,000 x G for 10 min.

In the wild type versus subGATA pull-down (Figure 4.3), bead-DNA complexes were incubated with approximately 300 μg (100 μl) of pre-cleared zebrafish IEC nuclear

extract, EMSA binding buffer (10mM Tris-Cl, 50 mM KCl, 1 mM DTT, pH 7.5), 1x Complete protease inhibitors, and 100 ng/μl Poly (dI-dC) for 4 hours with gentle rotation at 4°C. Complexes were washed four times with 25 μl of EMSA binding buffer containing 100 ng/μl Poly (dI-dC). Complexes were either eluted with 2x sample buffer (0.1 M Tris-Cl pH 6.8, 2% SDS, 10%, Sucrose, 0.008% bromophenol blue, 0.24 M B-mercaptoethanol) for 10 min at 85°C or stepwise with 25 μl of increasing concentrations of NaCl in EMSA binding buffer. Washes and elutions were dialyzed against Buffer D (20 mM HEPES, 20% glycerol, 0.1 M KCl, 20 mM MgCl2, 0.2 mM EDTA, pH 7.9) for 1hr at 25°C using 0.025 μm VSWP MF-Membrane Filters (Millipore VSWP02500) and used for EMSA and SDS-PAGE experiments.

In the wild type versus scrambled pull-down (Figure 4.4B), bead-DNA complexes were incubated with 100 μl (approximately 300μg) of pre-cleared zebrafish IEC nuclear extract, EMSA binding buffer (10 mM Tris-Cl, 50 mM KCl, 1 mM DTT, pH 7.5), 1x Complete protease inhibitors, and 100 ng/μl Poly (dI-dC) for 4 hours at 4°C in low protein binding tubes (Costar 1.7 ml pre-lubricated, 3207). In Figure 4.3B,C, complexes were washed four times with 250 μl EMSA binding buffer. Complexes were eluted stepwise with 25 μl of increasing concentrations of NaCl in EMSA binding buffer. Elutions were dialyzed against Buffer D (20 mM HEPES, 20 % glycerol, 0.1 M KCl, 20 mM MgCl2, 0.2 mM EDTA, pH 7.9) for 45 min. at 25°C using 0.025 μm VSWP MF-Membrane Filters (Millipore VSWP02500) and used for EMSA and SDS-PAGE experiments.

In Figure 4.4E-H, Bead-DNA complexes were incubated with 100 μl (approximately 300 μg) of pre-cleared zebrafish IEC nuclear extract, EMSA binding buffer (10 mM Tris-Cl, 50 mM KCl, 1 mM DTT, pH 7.5), 1x Complete protease inhibitors, 20μg/ml BSA, 100 ng/μl Poly (dI-dC), and 100 μg of unlabeled scrambled competitor

DNA for 4 hours at 4°C in low protein binding tubes (Costar 1.7ml pre-lubricated, 3207). Complexes were washed twice with 500 μl of 1x EMSA binding buffer containing 10 μg/ml scrambled competitor and 20 μg/ml BSA, twice with 500 μl of 1x EMSA binding buffer containing 20 μg/ml BSA, and twice with 500 μl 1x EMSA binding buffer. Complexes were eluted in one-step of 50 μl of 1 M NaCl in EMSA binding buffer for 3 minutes followed by a second denaturing elution with 2x sample buffer (0.1 M Tris-Cl pH 6.8, 2% SDS, 10%, Sucrose, 0.008% bromophenol blue, 0.24 M B-mercaptoethanol) for 10 min at 85°C.

For the pull-downs used in direct identification of peptides in complex mixtures (Table 4.1, 4.2), bead-DNA complexes were incubated with 100 μl (approximately 300 μg) of pre-cleared zebrafish IEC nuclear extract, EMSA binding buffer (10 mM Tris-Cl, 50 mM KCl, 1 mM DTT, pH 7.5), 1x Complete protease inhibitors, 100 ng/μl Poly (dI-dC), and 100 μg of unlabeled scrambled competitor DNA for 4 hours at 4°C in low protein binding tubes (Costar 1.7ml pre-lubricated, 3207). Complexes were washed six times with 500 μl of 1x EMSA binding buffer containing, 0.01% Triton X-100, 50 mM NaCl, and 100 ng/μl Poly (dI-dC). Bead-DNA-protein complexes were submitted directly to the UNC proteomics core for identification of peptides

For the competitor elution pull-downs (Figure 4.5), binding reactions were pre-incubated with 100 μl (approximately 300 μg) of pre-cleared zebrafish IEC nuclear extract, EMSA binding buffer (10 mM Tris-Cl, 50 mM KCl, 1 mM DTT, pH 7.5), 1x Complete protease inhibitors, and 100 μg of unlabeled scrambled competitor DNA in low protein binding tubes for 5 minutes on ice and then incubated with wild type DNA-bead complexes for 4 hours at 4°C. Complexes were washed (500 μl each wash) twice with 1x EMSA binding buffer containing 50 μg of unlabeled scrambled competitor and 50 μl of Nuclear Extract buffer (ActivMotif), twice with 500 μl of EMSA binding buffer containing

25 μg poly (dI/dC) and 50 μl of Nuclear Extract buffer (ActivMotif), and twice with EMSA binding buffer containing 50 μl of Nuclear Extract buffer (ActivMotif). Proteins were eluted with either 50 μl of 1x EMSA binding buffer containing 25 μg wild type unlabeled DNA or 25 μg scrambled unlabeled DNA.

4.5.9 Protein gel electrophoresis, staining, and mass spectrometry

Protein samples were denatured in 2x sample buffer (0.1 M Tris-Cl pH 6.8, 2% SDS, 10%, Sucrose, 0.008% bromophenol blue, 0.24 M B-mercaptoethanol), separated with NuPAGE 4-12% Bis Tris Gel (Invitrogen, NP0321PK2), and stained with either SilverQuest silver stain (Invitrogen, LC6070) or SyproRuby (Invitrogen, S-11791). Candidate bands were excised and in-gel digested with trypsin by the UNC Michael Hooker Proteomics Center and peptides identified using MALDI-TOF mass spectrometry. For identification of proteins in mixtures, proteins were digested in-solution and analyzed by LC-MALDI-TOF/TOF. Wild type and SubGATA bead-DNA-protein complexes were submitted directly to the UNC Michael Hooker Proteomics Center. Each sample was reduced, alkylated, and digested with trypsin. The peptides were extracted, lyophilized, and analyzed by LC-MALDI-TOF/TOF using a 90 min. gradient. Peptides were searched against the Uniprotein database (all species) and the International protein index (IPI) database (zebrafish).

SuGATA site-directed mutagenesis:

Name	Forward	Reverse	Cloning Method
SubGATA	TGTTCAAGGTCCAGTGTTTGAATTCAGGATTTAGTACACCATTAAACAAT	ATTGTTTAATGGTGTACTAAATCCTGAATTCAAACACTGGACCTTGAACA	Circular PCR/DpnI/EcoRI

Probes for EMSA and DNA Affinity Chromatography:

Name	Forward	Reverse	Modifications
in3.4-CR_EMSA	ACATGTTCAAGGTCCAGTGTTTGAGATAAGGATTTAGTACACCATTAA	TTAATGGTGTACTAAATCCTTATCTCAAACACTGGACCTTGAACATGT	
in3.4-CR_EMSA	ACATGTTCAAGGTCCAGTGTTTGAGATAAGGATTTAGTACACCATTAA		5' Biotin
SubGATA_EMSA	ACATGTTCAAGGTCCAGTGTTTGAATTCAGGATTTAGTACACCATTAA	TTAATGGTGTACTAAATCCTGAATTCAAACACTGGACCTTGAACATGT	
SubGATA_EMSA_Bio	ACATGTTCAAGGTCCAGTGTTTGAATTCAGGATTTAGTACACCATTAA		5' Biotin
Sub4_EMSA	ACATTAATCGATGACAGTGTTTGAGATAAGGATTTAGTACACCATTAA	TTAATGGTGTACTAAATCCTTATCTCAAACACTGTCATCGATTAATGT	
Sub4_EMSA_Bio	ACATTAATCGATGACAGTGTTTGAGATAAGGATTTAGTACACCATTAA		5' Biotin
Sub5_EMSA	ACATGTTCAAGGTCACATCGATTCGATAAGGATTTAGTACACCATTAA	TTAATGGTGTACTAAATCCTTATCGAATCGATGTGACCTTGAACATGT	
Sub5_EMSA_Bio	ACATGTTCAAGGTCACATCGATTCGATAAGGATTTAGTACACCATTAA		5' Biotin
Sub6_EMSA	ACATGTTCAAGGTCCAGTGTTTGCAATCGATGGTTAGTACACCATTAA	TTAATGGTGTACTAACCATCGATTGCAAACACTGGACCTTGAACATGT	
Sub6_EMSA_Bio	ACATGTTCAAGGTCCAGTGTTTGCAATCGATGGTTAGTACACCATTAA		5' Biotin
Sub7_EMSA	ACATGTTCAAGGTCCAGTGTTTGAGATAAGGATGCATCGATAAATTAA	TTAATTTATCGATGCATCCTTATCTCAAACACTGGACCTTGAACATGT	
Sub7_EMSA_Bio	ACATGTTCAAGGTCCAGTGTTTGAGATAAGGATGCATCGATAAATTAA		5' Biotin
Scrambled_EMSA	CTCAATTGAGCATTAGGTGCGAATTCCATTGTTAAGAAAAAGTATTCG	CGAATACTTTTTCTTAACAATGGAATTCGCACCTAATGCTCAATTGAG	
Scrambled_Bio	CTCAATTGAGCATTAGGTGCGAATTCCATTGTTAAGAAAAAGTATTCG		5' Biotin

Table 4.3: Primers and oligo sequences used in this study

4.6 Acknowledgements

I am grateful to David Smalley and Nedyalka Dicheva from the UNC Proteomics Core and Ben Major for their assistance with mass spectrometry; to Bill Marzluff for help designing EMSA experiments; to Dylan Taatjik for advice on DNA chromatography; and to the labs.

CHAPTER 5

A Pilot Atlas of Open Chromatin in the Intestinal Epithelium of Mouse and Zebrafish

5.1 Overview

The body surfaces of humans and other animals are colonized at birth by microorganisms. The majority of these microbial residents exist within gastrointestinal tract (GI) communities, where they engage in complex symbioses with host epithelial cells. The host genome encodes the ability to respond to microbial stimuli making it the nexus and historical record for this ancient symbiosis. Exploration into the molecular mechanisms mediating host-microbiota symbiosis and dysbiosis has largely focused on the protein-coding portion of the host genome. However, non-genic functional DNA regions govern transcriptional programs that pattern organism development, shape cell identity, generate phenotypic diversity, maintain homeostasis, and drive disease progression. In this light, I have designed and implemented experiments in mouse and zebrafish that aim to uncover the non-genic regions of the vertebrate genome that mediate host intestinal epithelial cell (IEC) response to the microbiota. I applied DNase-seq to IECs isolated from the duodenum, ileum, and colon of germ-free (GF) or conventionally-raised (CONV-R) mice to generate open chromatin maps in each condition. To probe the evolution of host transcriptional regulation in the intestine, I generated open chromatin maps using FAIRE-seq on dissected gastrointestinal (GI) tracts from GF and CONV-R zebrafish larvae as well as adult zebrafish IECs. To demonstrate the utility of this information, I present preliminary analysis of pilot results

from DNase-seq data from CONV-R mouse ileal IECs and FAIRE-seq data from CONV-R adult zebrafish IECs. These experiments should begin to elucidate the mechanisms mediating over 450 million years of co-existence and co-evolution of vertebrate hosts with their intestinal microbiota.

5.2 Introduction

From birth until death, the GI tracts of all animals are home to vast communities of microorganisms functioning as a non-self metabolic organ broadly shaping the physiologic potential and fitness of the host organism [37,294]. The intestinal epithelium directly interfaces with the microbiota engaging in an ancient symbiosis where homeostatic balance requires proper control of gene expression in space and time within the epithelial layer. This is highlighted by the observation that aberrant gene regulation in the mammalian intestine can lead to pathological conditions such as inflammatory bowel disease [295], obesity [11], and cancer [296]. A more complete understanding of the mechanisms mediating host response to microbial activity within the GI tract is needed if therapeutic manipulations of the microbiota and the host responses they evoke are to be achieved.

Genomes include a historical record of host-microbiota symbiosis [24]. The host genome encodes transcriptional programs that enable specification, differentiation, and function of each cell type within the body. The genome also encodes the transcriptional plasticity for cells to respond to diverse stimuli deriving from other host cells or the environment. Advances in genomics are rapidly uncovering the genic and non-genic functional DNA in many cell types from diverse organisms [23,67,68,69 2012,70]. This work has expanded on the historical theory [297] and its modern realization [222] that networks of modular *cis*-regulatory DNA (defined here as *cis*-regulatory modules, CRMs) govern cell-type specific transcriptional programs, and allows discovery of CRMs

genome-wide. Two complementary genome-wide methods, DNase-seq [136,137] and FAIRE-seq [138,139], take advantage of the observation that eviction or destabilization of nucleosomes from chromatin is a characteristic feature of functional CRMs in eukaryotic genomes. DNase-seq is the genome-wide extension of the classical DNase I footprinting assay [140]. DNase I footprinting harnesses the feature that protein factors binding naked DNA block DNase I mediated enzymatic cleavage of underlying nucleotides, thus giving a quantitative footprint of the DNA binding factor. In the context of chromatin, the vast majority of DNA is protected from digestion by nucleosomes whereas regions adjacent to transcription factor binding are accessible or hypersensitive to DNase I cleavage. This allows identification of “open” chromatin regions, which have very strong correlations with a variety of other markers (transcription factor binding, histone marks) of active non-coding regulatory function [136]. Furthermore, within the “open” region defined by increased DNase I sensitivity there is often a discernible footprint of transcription factors bound to their cognate DNA sequence that is distinguished by a local decrease in DNase I sensitivity [142]. Combined with a high signal-to-noise ratio, DNase-seq offers a powerful and validated method to discern nucleosome depleted regions as well as transcription factor-DNA interactions across the genome.

Formaldehyde-Assisted-Isolation-of-Regulatory-Elements (FAIRE) is an alternative approach to discover “open” chromatin based on differences in cross-linking efficiencies between DNA and nucleosomes compared to DNA and sequence-specific DNA-binding proteins. In this assay, cells are covalently cross-linked briefly with formaldehyde, lysed and sonicated, and sheared chromatin is extracted with phenol/chloroform. Extraction enriches unbound DNA into the aqueous phase and protein-bound DNA is trapped to the organic/aqueous phase interface. Unbound DNA is isolated and assayed for locus-specific (via quantitative PCR) or genome-wide (via

microarray or high-throughput sequencing) enrichment patterns. The average signal-to-noise ratio for FAIRE-seq is not as high as DNase-seq, and it has yet to be proven as a method for elucidating transcription factor footprints [139,141]. However, FAIRE does not require nuclei isolation so samples do not need to be in single cell suspensions, and other experimental practicalities [139] position FAIRE-seq to be amenable to higher-throughput capabilities. Both genomics tools, DNase-seq and FAIRE-seq, can uncover a range of cell-type specific elements (promoters, enhancers, silencers, insulators, locus control regions) and do not require an antibody [141]. The impact of environmental factors, such as changes in microbial community composition and diet, on open chromatin dynamics is not well known, and neither assay has been applied to primary intestinal epithelial cells in mouse or zebrafish.

The mouse and zebrafish gnotobiotic models present a unique opportunity to understand the effect of host-microbe symbiosis on *cis*-regulatory evolution. It has been established that the microbiota impacts gene regulation on a genomic scale [87] and a variety of conserved responses to the microbiota are known [88-90,220]. Intestinal epithelial architecture and function is well conserved across vertebrate lineages [45,47], and intestinal epithelial cells (IECs) are relatively simple to isolate from primary tissues representing an abundant source material for genomics based chromatin assays. Furthermore, methods for downstream functional analysis of CRMs *in vivo* are available for both organisms [290,298]. Here I present methods for, and pilot results from, a genome-wide atlas of open chromatin in intestinal epithelial cells from mouse and zebrafish. I provide preliminary analyses to exemplify how these datasets can be used to understand intestine-specific and microbial-responsive gene expression. I have generated multiple biological replicates of DNase-seq or FAIRE-seq samples from mouse or zebrafish, respectively, with and without a microbiota. Most of these samples are currently in the sequencing pipeline, and I discuss future plans to analyze the

anticipated data. This work constitutes an important leap toward elucidating the history of host-microbe co-evolution.

5.3 Results

5.3.1 Strategy to discover microbially-responsive CRMs genome-wide

I designed and implemented multiple strategies to elucidate open chromatin in germ-free (GF) and conventionally-raised (CONV-R) mice and zebrafish (Figure 5.1, 5.2). Primary mouse intestinal epithelial cells (mIECs) were isolated from the duodenum, ileum, and colon from three GF and three CONV-R C57BL/6 mice at approximately 10 weeks of age and processed for DNase-seq and RNA extraction (Figure 5.1A). The C57BL/6 background was chosen because there are extensive comparative datasets probing the gene expression responses to the microbiota [87] and datasets describing the chromatin landscapes in other tissues {Shen, 2012}. In the zebrafish, I performed FAIRE-seq on two sets of 100 pooled dissected GI tracts from GF and CONV-R 6dpf zebrafish (Figure 5.1B). This time point was chosen because we have substantial knowledge concerning the effect of the microbiota on zebrafish physiology at 6dpf and also the practical consideration that it is not currently feasible to rear GF zebrafish to adults [90,94,109,191,195,290,299]. Because the GI tract dissections are not IEC specific, I also wanted to generate data on CONV-R adult zebrafish IECs (zIECs) using FAIRE-seq and/or DNase-seq in order to compare open chromatin profiles with 6dpf zebrafish and mouse mIECs (Figure 5.2). There are no established methods for either DNase-seq or

FAIRE-seq in the mouse or zebrafish intestinal epithelium, and I will therefore discuss my efforts to establish these methods in our lab in the following sections.

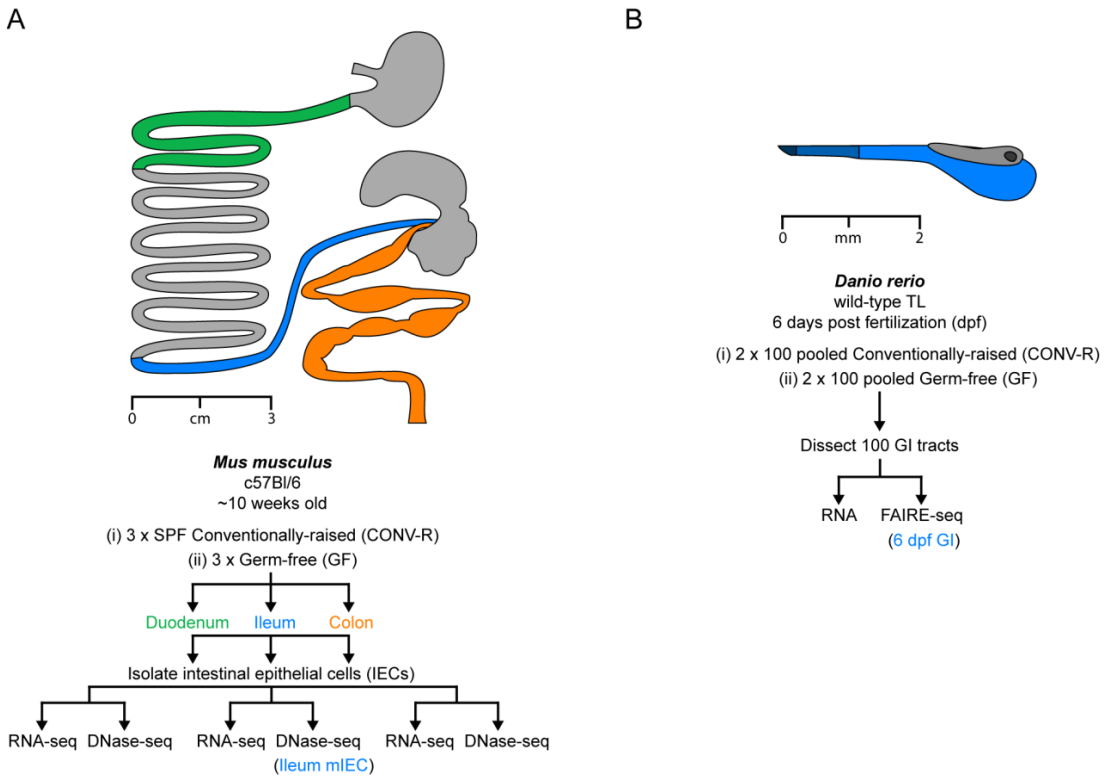


Figure 5.1: Experimental strategy to discover microbiota regulated CRMs
Experimental outline shows the samples and current pilot datasets generated in this study. (A) Schematic of the mouse GI tract showing the stomach (dark gray), duodenum (green), jejunum (light blue), ileum (blue), cecum (light gray), and colon (orange). The GI tract is loosely drawn to scale and based off of Hume, 1995. Mouse intestinal epithelial cells (mIECs) from the duodenum, ileum, and colon of three sibling 10 week old C57BL/6 mice reared since birth with a specific pathogen free (CONV-R) microbiota or reared under axenic conditions (GF) were harvested and prepared for DNase-seq and RNA-seq. Currently the results for DNase-seq of 1 pilot CONV-R SPF mice are available and it is referred to as ileum mIEC. (B) Two sets of GI tracts dissected from one hundred 6dpf zebrafish unfed and reared with a conventional (CONV-R) zebrafish microbiota or reared GF were harvested and prepared for FAIRE-seq. GI tract dissections include cells from the exocrine (light gray) and endocrine pancreas (dark grey) and the and the anterior (light blue), middle (blue), and posterior intestine (dark blue), but excludes swim bladder and liver. Currently the sequencing results from 1 pilot CONV-R sample are available. Drawn to approximate scale.

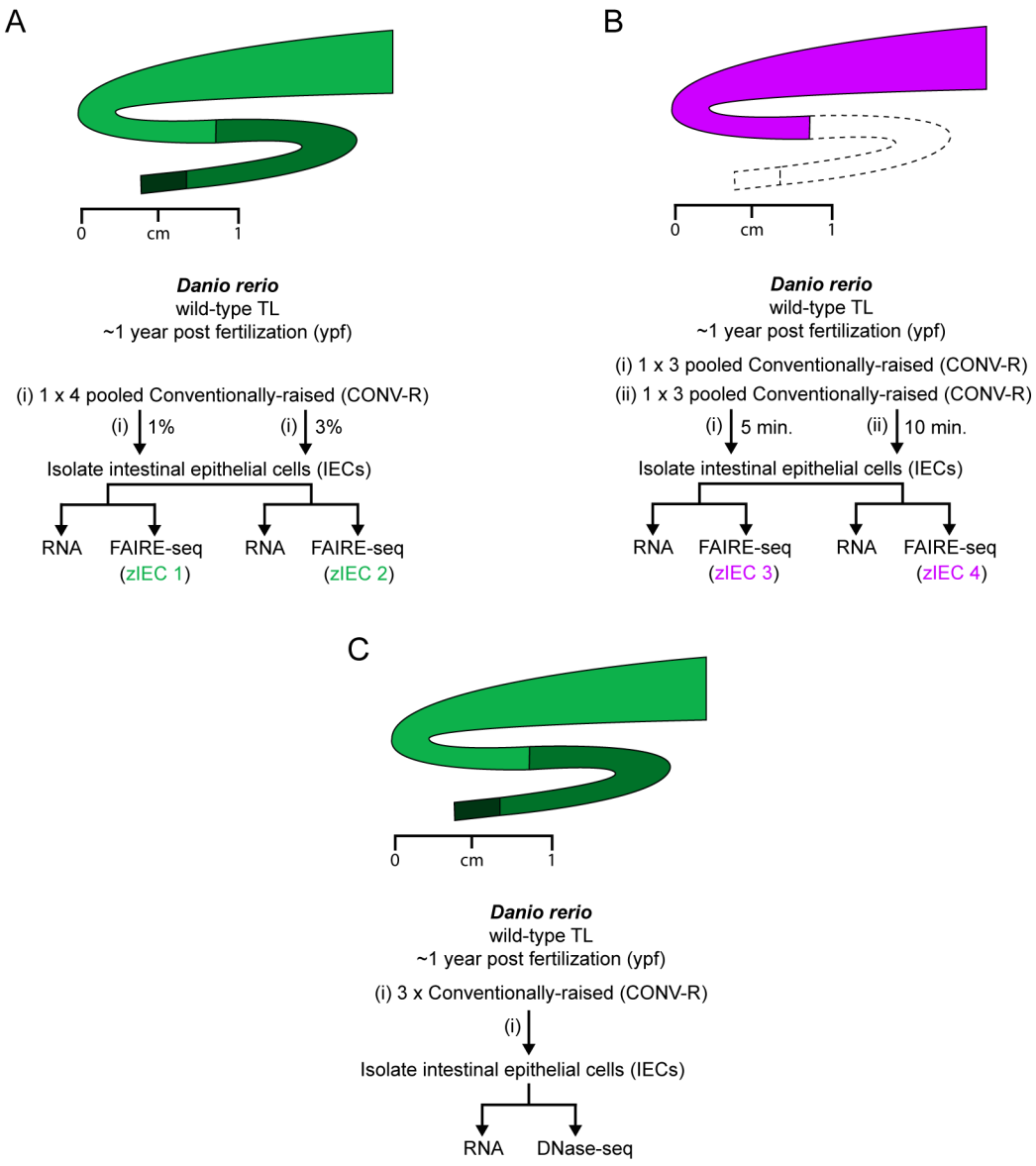


Figure 5.2: Experimental strategy and description of zebrafish IEC datasets
(A) IECs isolated from whole intestines from four adult zebrafish fed and reared under conventional conditions were prepared for FAIRE-seq. IECs from the four fish were first pooled, then split into two FAIRE conditions (1% formaldehyde, zIEC 1; or 3 % Formaldehyde, IEC 2) in order to optimize the FAIRE protocol for these cell types. (B) Two sets of IECs isolated from segment 1 of three adult zebrafish fed and reared under conventional conditions were prepared for FAIRE-seq. One set was treated with 1% formaldehyde for 5 minutes and the other set treated with 1% formaldehyde for 10 minutes. (C) Biological replicates of IECs isolated from whole intestines of one adult zebrafish fed and reared under conventional conditions were prepared for DNase-seq. Note that cells were reserved for RNA isolation from all zebrafish intestinal epithelial cells samples. Drawn to approximately to scale.

5.3.2 Establishing DNase-seq in the mouse intestinal epithelium

Cell types can often display specific patterns of chromatin accessibility [67,141], therefore it is often desirable to minimize the cell type complexity in the sample when using methods such as DNase-seq and FAIRE-seq. The small and large intestine are tissues composed of a myriad of differentiated cell types. In order to enrich for epithelial cells from diverse regions of the murine GI tract, I sloughed the epithelial layer from the duodenum, ileum, and colon (Figure 5.3A) with 30 mM EDTA using established protocols conducive to post-isolation cell culture [291]. Flow cytometry of the mIEC cell preparation from the entire small intestine revealed that over 90% of cells stained positive for a brush border marker (EpCAM) and negative for a marker of endothelial cells and blood cells (CD31) (Figure 5.3B) suggesting that most cells in the preparation are epithelial derived. I followed closely the Song et al. protocol [137] from the lab of Greg Crawford in my first attempts to perform DHS on mIECS from the duodenum, ileum, and colon. I first tested various concentrations of Igepal detergent (0% - 0.2%) and discovered promisingly that cell lysis with 0.1% Igepal sufficiently disrupted the plasma membrane leaving the nuclei intact (Figure 5.3C). The detergent treatment was able to dissociate many of the large epithelial sheets into smaller sheets and single cell nuclei suspensions, however it should be noted that the resultant DNase digestions were a mixture of epithelial sheets and single cells. Trypan blue staining revealed nearly 100% cell lysis using 0.1% Igepal (data not shown). In contrast to the Crawford lab protocol, incubation of nuclei isolated from the duodenum, ileum, and colon with increasing concentrations of exogenous DNase I for 15 minutes resulted in digestion of high molecular weight (HMW) DNA even when no exogenous DNA was added (Figure 5.3D). Intestinal epithelial cells have been shown to harbor active endongenous DNases at high concentrations [280]. I hypothesized that endogenous DNase activity could be harnessed to digest DNA at open chromatin regions over a time course. Indeed,

incubation of mIECs from the duodenum, ileum, and colon resulted in DNase digestion patterns similar to those observed in published protocols (Figure 5.3E). In collaboration with Chris Frank in the lab of Greg Crawford, we were successful in generating DNase-seq libraries from duodenal, ileal, and colonic mIECs (Figure 5.3F,G). We first sequenced one pilot library prepared from the CONV-R ileum at the Duke University Genome Sequencing and Analysis Core Resource and aligned 156,832,519 reads to the mouse genome (mm9) yielding 143,249,947 usable alignments. Raw aligned sequencing reads were smoothened using a kernel density estimation function called Parzen windowing [300,301], which allows identification of DNase hypersensitivity sites (DHSs) from uniquely mapped tags. The number of reads and mapping data can be found in Table 5.4. Visualization of raw reads and DH sites in the UCSC genome browser revealed striking digestion patterns with signal-to- noise ratios visually comparable to published DNase-seq datasets (Figure 5.3H). Regulatory regions at the *Villin-1* locus have been characterized *in vivo* [298] and used extensively as IEC-specific drivers in the mouse. Inspection of this locus revealed two strong DH sites within 500 bases of the transcription start site (TSS) as well as multiple peaks within the 1st intron. Both sets of regions were previously shown to be required for proper expression of reporter constructs in the intestinal epithelium [298]. DNase-seq identified a third peak within intron 2 and the function of this novel DH site is currently unknown. Inspection of many other loci (data not shown) strongly suggested that using endogenous DNase activity could aptly capture the open chromatin state of IECs in the mouse ileum. I therefore proceeded with library preparation and submitted samples from three biological replicates from GF and CONV-R duodenum, ileum, and colon for sequencing. At the submission of this dissertation, we are awaiting sequencing results from these samples.

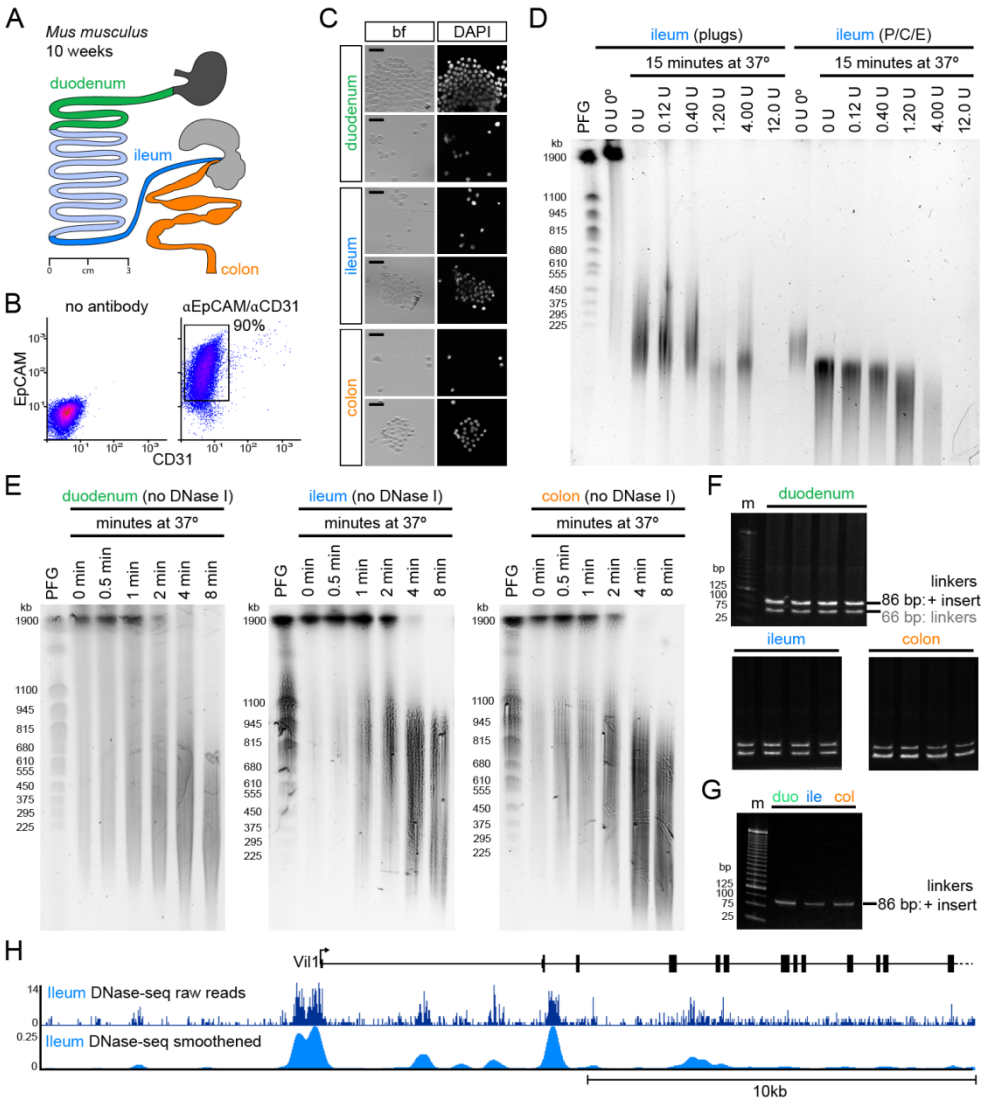


Figure 5.3: Establishment of DNase-seq in mouse IECs

(A) Schematic of portions of the mouse GI tract showing the stomach (dark gray), duodenum (green), jejunum (light blue), ileum (blue), cecum (light gray), and colon (orange). The GI tract is loosely drawn to scale and based off of Hume, 1995. (B) Fluorescence activated cell sorting of primary midgut (duodenum, jejunum, ileum) IECs labeled with antibodies marking either brush border cells (EpCAM) or endothelial cells/leukocytes/platelets (CD31). Approximately 90% of isolated cells were EpCAM positive and CD31 negative suggesting epithelial origin. (C) IECs lysed with 0.1% Igepal from the duodenum, ileum, or colon stained with DAPI shows intact nuclei. Note that many cells are fully lysed and the nuclei float free, however there are some intact epithelial layers in each preparation. 90% of cells stained positive for trypan blue (not shown) (D) Pulse-field gels from two experiments where ileal cell preparations where incubated on ice (0°C) or with increasing concentrations of exogenous DNase I (U = Units of DNase I) for 15 minutes at 37°C. Reactions were terminated using either EDTA and agarose plugs (plugs) or using lysis buffer followed by phenol-chloroform extraction (P/C/E). High molecular weight (HMW) DNA is stable at 0°C and in EDTA soaked plugs however even with no addition of exogenous DNase I there was significant digestion of DNA when incubated at 37°C. Note that phenol-chloroform extraction followed by ethanol precipitation results in no HMW DNA. PFG = pulse-field gel yeast chromosome marker. (E) Pulse-field gel showing that endogenous DNases begin digesting DNA as soon as 0.5 minutes after moving nuclei to 37°C and by 8 minutes most HMW DNA is digested. This was consistent for duodenum, ileum, and colon. The observed

digestion pattern is similar to optimal digestion patterns in [137] and we proceeded with making libraries from the 2, 4, and 8 minutes digestions that used endogenous DNase activity. (F-G) Polyacrylamide gel electrophoresis showing that DNase-seq libraries were successfully made for the duodenum, ileum, and colon (86 base-pairs). (H) Raw and smoothened DNase-seq reads from CONV-R ileum at the Vil1 locus. Note strong peaks at the transcription start site, as well as peaks within the first intron, both of which are required for IEC-specific expression [298].

5.3.3 Establishing DNase-seq in the zebrafish intestinal epithelium

Due to the apparent success of using endogenous DNases to elucidate the open chromatin landscape in the mouse ileum, I attempted a similar protocol in adult zebrafish intestinal epithelial cells (zIECs) (Figure 5.4A). Similar to mouse IECs, cell lysis with 0.1% Igepal disrupted the plasma membrane leaving nuclei intact (Figure 5.4B). Zebrafish are poikilothermic teleosts where the surrounding water temperature generally maintains body temperatures, which in our lab is 25-28°C. I reasoned that endogenous DNase activity should therefore be high even at 25°C. Indeed, incubation over a time course from 0 to 8 minutes resulted in a corresponding increase in HMW DNA digestion (Figure 5.4C). However, all HMW DNA was digested even when nuclei were left on ice (Figure 5.4C). It is possible that during the sloughing procedure DNases could become activated and would exert their function during the 5-10 minutes it takes to slough the epithelium at 25°C. I therefore performed all steps in the DNase protocol in the 4°C cold room and this resulted in significantly less, though not optimal, cleavage at 0 minutes at 25°C with digestion increasing over time during the 25°C incubations (Figure 5.4C). I created sequencing libraries from these samples and we are currently awaiting sequencing results to determine if this method is able to distinguish open chromatin in the zebrafish intestinal epithelium.

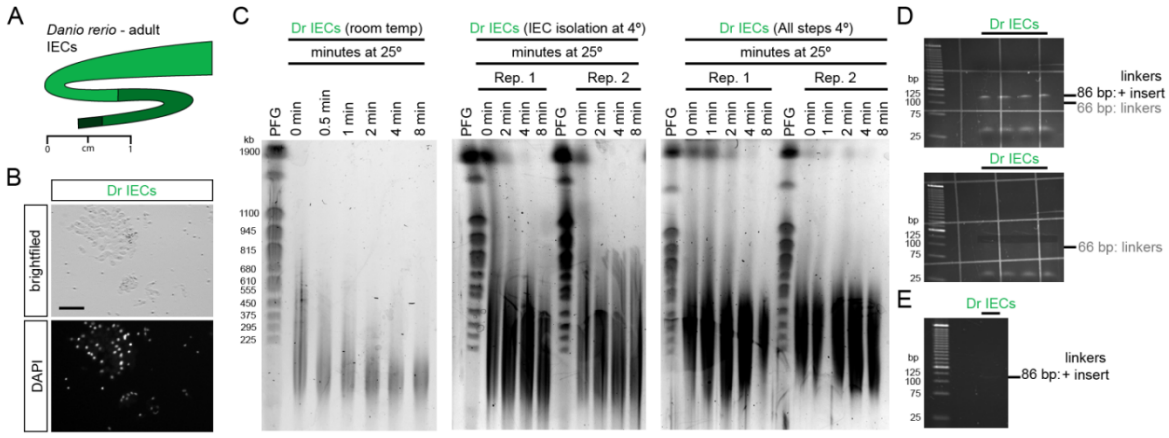


Figure 5.4: Establishment of DNase-seq in zebrafish IECs
(A) zIECs from the whole intestine of one to three adult zebrafish were used to make in DNase experiments. (B) IECs lysed with 0.1% Igepal are stained with DAPI to show intact nuclei. Note that many cells are fully lysed and the nuclei float free, however there are some intact epithelial layers in each preparation. Greater than 90% of cells stained positive for trypan blue (not shown). (C) Pulse-field gel stained with ethidium bromide of plugs from zebrafish IEC cell preparations that were incubated on ice (0 minutes at 25°C) or for increasing time periods at 25°C. The gel shows that all high molecular weight (HMW) DNA is digested by endogenous DNases even when left on ice (compare with mouse, Figure 5.3). PFG = pulse-field gel yeast chromosome marker. Note that in this experiment cells were sloughed from the epithelium at 25°C. (D) Pulse-field gel of plugs from two replicate zIEC cell preparations in which all steps post dissection are performed in the 4°C cold room. There is HMW DNA when the nuclei are left on ice (0 min at 25°C) and HMW DNA is digested over time. (E) All steps including the dissection were performed in the 4°C cold room. There appears to be no significant decrease in HMW digestion at 0 min compared with panel D. (F-G) Polyacrylamide gel electrophoresis showing that DNase-seq libraries were successfully made by combining plugs from 0 min., 2 min., 4 min. digestions (86 base-pairs). We are currently awaiting sequencing results.

5.3.4 Establishing FAIRE-seq in the zebrafish 6dpf GI tract and adult IECs

FAIRE can be performed on isolated cells and on intact tissues, and I set out to develop methods for FAIRE-seq using adult zIECs and 6dpf GI tracts (6dpf GI) (Figure 5.5A). I tried a variety of conditions in order to ascertain the optimal FAIRE conditions for these samples. I isolated IECs from the anterior intestine (samples zIEC 3 and zIEC 4) or from the entire intestinal tract (zIEC 1 and zIEC 2). For zIEC 3 and zIEC 4, I incubated segment 1 cells with 1% old (2+ years) formaldehyde solution for either 5 or 10 minutes at room temperature, respectively. For zIEC 1 and zIEC 2, I incubated cells with either 1% or 3% new formaldehyde solution for 5 minutes at room temperature. For 6 dpf GI

tracts, I dissected 100 tracts in pools of 25 and incubated them with 4% old formaldehyde solution for 10 minutes at room temperature (Figure 5.5A). For all samples I quenched the formaldehyde with 125 mM glycine for 5 minutes at room temperature. I sonicated each sample until the desired amount of sonication (~300 bp fragments, see Methods) was achieved. Importantly, no DNA laddering or shearing was observed in un-sonicated samples for either 6dpf GI (Figure 5.5B) or zIECs (data not shown) suggesting that nucleases were no longer active post-fixation. Approximately 240 ng (zIEC 1), 60 ng (zIEC 2), 272 ng (zIEC 3), 241 ng (zIEC4), and 40 ng (6dpf GI) of FAIRE DNA was acquired per ml of sample (2 ml total) with a FAIRE/input ratio approximately 7.5% (zIEC 1), 2.2% (zIEC 2), 9% (zIEC 3), 11% (zIEC 4), and 2.7% (6dpf GI) (Figure 5.5C). All samples showed enrichment at previously defined regulatory regions at the *angptl4* locus relative to a randomly selected region downstream from *angptl4* using quantitative PCR (Figure 5.5D). Sequencing libraries were verified using an Agilent Bioanalyzer and submitted for sequencing to the UNC High-Throughput Sequencing Core Facility. The number of reads and mapping data can be found in Table 5.4. Strikingly, inspection of the *angptl4* locus revealed strong peaks in all samples at the TSS and the 3' portion of intron 3 with peaks overlapping both the islet (in3.3) and intestinal (in3.4) CRM (Figure 5.5F). The signal-to-noise ratio varied across samples and loci, however the zIEC 2 sample consistently had the highest signal-to-noise ratio as predicted by qPCR (Figure 5.5D) and the FAIRE/input ratios (Figure 5.5C). Cumulatively, these data reveal successful establishment of methods for elucidation of open chromatin in zIECs and 6dpf GI tracts.

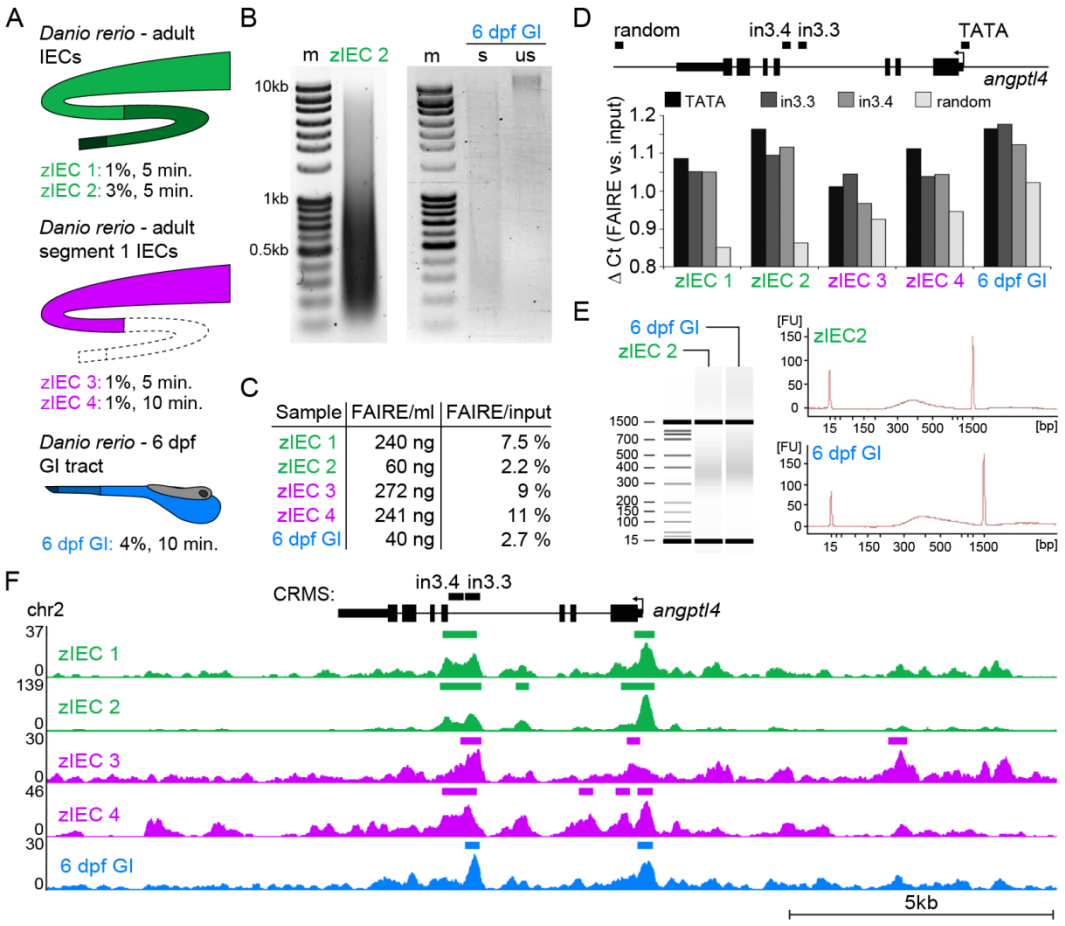


Figure 5.5: Establishment of FAIRE-seq in zebrafish 6dpf GI tracts and adult IECs
(A) Schematic of the samples used to create FAIRE-seq libraries. zIECs from adult whole intestines (green, zIEC 1 and zIEC 2) or segment 1 only (purple, zIEC 3 and zIEC 4) were isolated and processed for FAIRE-seq. (B) Representative images show ethidium bromide stained gels of sonicated (s) DNA from adult zIECs (10 cycles) and 6dpf GI tracts (9 cycles). Note that the number of cycles required for effective sonication varied from 6-13. Unsonicated (us) DNA from 6dpf GI tracts is shown and similar results were obtained for adult zIECS (not shown). (C) Table shows the amount of FAIRE DNA isolated from 1ml of each sample and the ratio of FAIRE DNA to input DNA in which the cross-links were reversed. (D) Quantitative PCR on select regions at the angptl4 locus were assayed for FAIRE enrichment relative to input DNA. There is enrichment in each library in regulatory regions previously described in Chapter 3 (in3.4 is the intestinal CRM, in3.3 is the islet CRM, TATA refers to the TATA box at the proximal promoter, and random is region downstream of *angptl4* chosen randomly). Note that zIEC 2 has both the lowest FAIRE/input ratio and highest delta Ct. (E) FAIRE-seq library preparations were analyzed for quality using an Agilent Bioanalyzer and digital gels and histograms from two representative libraries are shown. Peaks should be approximately 300-500 base pairs and of sufficient concentration. (F) FAIRE-seq results at the *angptl4* locus for each sample. Note that both the region around the transcription start site and the 3' portion of intron 3 are enriched in all samples. The signal-to-noise ratio is particularly good in the zIEC2 sample (see y-axis). Peak calls (see Methods) are shown as bars above each track.

5.3.5 Preliminary analysis of pilot DNase-seq and FAIRE-seq CONVR datasets

We expect to have at least ten DNase-seq datasets from the mouse intestinal epithelium, two DNase-seq datasets from zIECs, four FAIRE-seq datasets from zIECs, four FAIRE-seq datasets from zebrafish 6dpf GI tracts, and corresponding RNA expression from each dataset. This represents a rich resource requiring extensive multilayered analyses of which this chapter is a preliminary introduction. The purpose of this results section is therefore to (i) provide preliminary analysis of the ileum mIEC DNase-seq and zIEC FAIRE-seq data to show that these pilot data capture the regulatory landscape of the intestinal epithelium, and (ii) illustrate examples of how this data can be used. I will then discuss specific questions that more sophisticated analyses of multiple genome-wide datasets can help answer. Unpublished DNase-seq datasets from liver and kidney from C57/Bl6 mice were generated in the Crawford lab and were used in comparison with the ileum datasets with permission (G. Crawford, personal communication).

5.3.6 General features of DNase-seq in ileal mIECs

We analyzed the distribution of DH sites across the genome with respect to genes based on RefSeq gene annotations (Figure 5.6A). Out of the top 100,000 peaks (called using F-seq, [301]) 19% map to regions within 2kb upstream of an annotated transcription start site, whereas 34.3% map to intergenic regions. Approximately 36% of DH sites are located within introns and less than 10% map to exonic regions. However, looking specifically at the highest scoring 25% of DH sites reveals that over half map to regions within 2kb upstream of a gene supporting published reports describing the extreme hypersensitivity of promoter regions [136]. The weakest scoring DH sites are still significantly more susceptible to digestion than background (as defined by peak calling cutoffs) and their categorical distribution does not differ extensively from what is

seen for all sites. This data supports the authenticity of weaker scoring regions. The non-random distribution of peaks and strong hypersensitivity at promoter regions are consistent with published reports [136] further supporting the validity of this dataset.

We next asked if DNase hypersensitivity at promoter regions correlates with gene expression. Indeed, genes highly expressed in the ileum [87] have strong hypersensitive sites near the TSS, whereas genes lowly expressed in the ileum do not (Figure 5.6B). The Spearman correlation was moderate (0.59), but note that the gene expression dataset was from the entire ileum from mice reared in a different facility and not limited to IECs. Logically, an increased number of sequencing reads distinguish strong DH sites relative to weak DH sites. However, these reads are not randomly distributed within the hypersensitive site and can be used to identify protein-DNA interactions within a hypersensitive region [142,302]. Inspection of raw DNase-seq reads at a highly expressed gene in the ileum (*Villin 1*) revealed a depletion of DNase I cleavage sites overlapping a conserved TATA binding motif within the putative core proximal promoter (Figure 5.6D). These results suggest the utility of this and future replicate datasets for high-resolution *in vivo* footprinting across the intestinal epithelial genome.

Functional DNA elements often evolve under negative selection resulting in increased sequence conservation compared to non-functional DNA. I next examined the degree of conservation using PhastCons [303] scores across 1kb centered on each peak for the top 50,000 peaks from the ileum DNase-seq dataset (Figure 5.6E). This analysis revealed conservation scores well above background levels with the highest conservation located at the peak center. Similar conservation plots and scores were obtained for liver and kidney datasets (Figure 5.6E). Separating DH sites into genomic feature categories and re-analyzing conservation revealed substantially higher scores for promoter and exonic associated peaks compared to intergenic and intronic peaks,

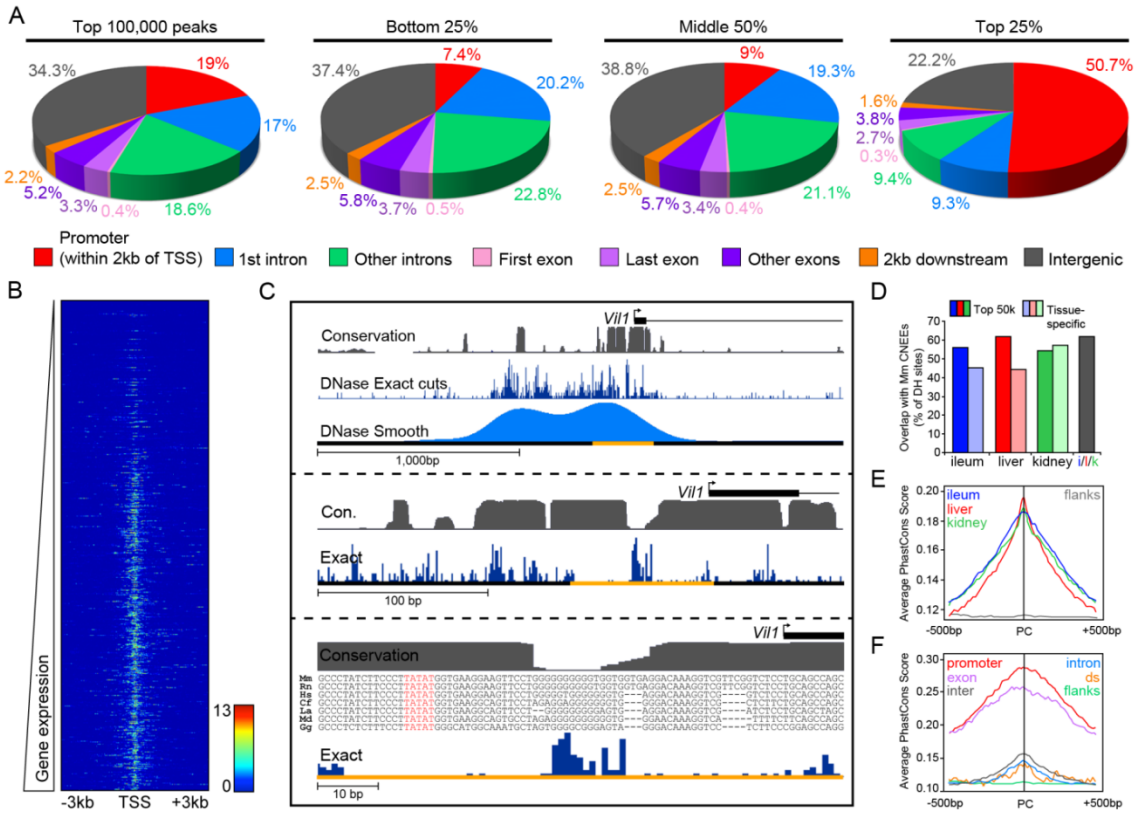


Figure 5.6: General features of DNase-seq open chromatin sites

(A) The locations of ileum DNase I hypersensitive sites relative to RefSeq gene annotations. Shown are the top 100,000 sites (Parzen scoring) broken down into bottom 25%, middle 50%, and top 25%. (B) Heat map of DNase-seq signals +/- 3kb around transcription start sites (TSSs) ordered by gene expression in the ileum [87]. Spearman correlation score is 0.59. (C) Conservation and DNase hypersensitivity footprint at the *Villin 1* gene promoter at three views of increasing resolution. Each peak in the exact cuts track represents the number of times that base appeared as the first base in the aligned read. Note that the highly conserved TATA box (red) is protected from DNase digestion revealing a discernible footprint. (D) Histogram showing the percentage of Top 50,000 (dark) or tissue-specific (light) DH sites from the ileum, liver, or kidney that overlap with mouse conserved non-exonic elements (CNEEs). Tissue-specific is defined as those DH sites from each tissue that do not intersect with the union of DH sites from the other two tissues. The gray bar labeled i/l/k represents the intersection of Top 50,000 DH sites from the ileum, liver, and kidney. See Figure 4.8A for more information. (E) Conservation plot showing the average PhastCons score of 500 bp surrounding the center of the Top 50,000 DH sites from the ileum (blue), liver (red), and kidney (green). Flanking DNA (gray) is defined here as 1000 bp flanking top 50,000 ileum peaks offset by 500 bp with no intersection with Refseq genic DNA. (E) Conservation plot showing average PhastCons score of the top 100,000 peaks overlapping each feature (within 2kb upstream of the TSS (promoter, red), all exons (purple), intergenic (inter, gray) all introns (blue), within 2kb downstream of the gene (ds, orange), and flanking DNA (flanks, green) used to define background levels).

however all sets of DH sites were more conserved than background sequences (Figure 5.6F). Notably, more than half of the Top 50,000 DH sites from the ileum, liver, and kidney overlapped with conserved non-exonic elements (CNEEs) [120] in the mouse genome. Similar overlap was observed when the Top 50,000 DH sites were filtered for tissue-specificity (Figure 5.6D). Taken together, the pilot DNase-seq dataset from mouse ileal IECs exhibits hallmark features of other datasets that use exogenous DNase I activity to map the *cis*-regulatory landscape genome-wide.

5.3.7 Dnase-seq elucidates putative cell-type specific CRMs

The ileal intestinal epithelium is composed of at least four principal differentiated cell types: absorptive enterocytes, enteroendocrine cells, goblet cells, and paneth cells (Figure 5.7A,B). I inspected the DNase-seq landscape at the loci of multiple genes that have known cell-type specific expression patterns in the intestinal epithelium. Both *Intestinal alkaline phosphatase* (*Alpi,* Figure 5.7C) and *Intestinal fatty acid binding protein* (*Fabp2*, not shown) are common markers of differentiated enterocytes [304,305] and have distinct DH sites not present in DNase-seq datasets from the liver or kidney. *Cholecystokinin* (*Cck*, Figure 5.7C) and *Chromogranin A* (*Chga*, not shown) are expressed by enteroendocrine cells and have common and distinct DH peaks compared to liver and kidney datasets [306,307]. *Mucin 2* gene expression is characteristic of goblet cells and the *Muc2* locus has multiple DH sites not present in the liver or kidney [308] (Figure 5.7C). Finally, *Regenerating islet-derived protein 3 gamma* (*Reg3*γ; Figure 5.7C) and *Lysozyme 1* (*Lyz1*, not shown) have strong expression in paneth cells and have corresponding ileal-specific open chromatin regions [154]. The non-overlap of ileum and liver or kidney peaks at these genes is not due to poor quality datasets from the liver and kidney. For example, ubiquitously expressed genes such as ß-actin have very strong DH peaks covering the same regions at the ß-actin locus in ileum, liver, and

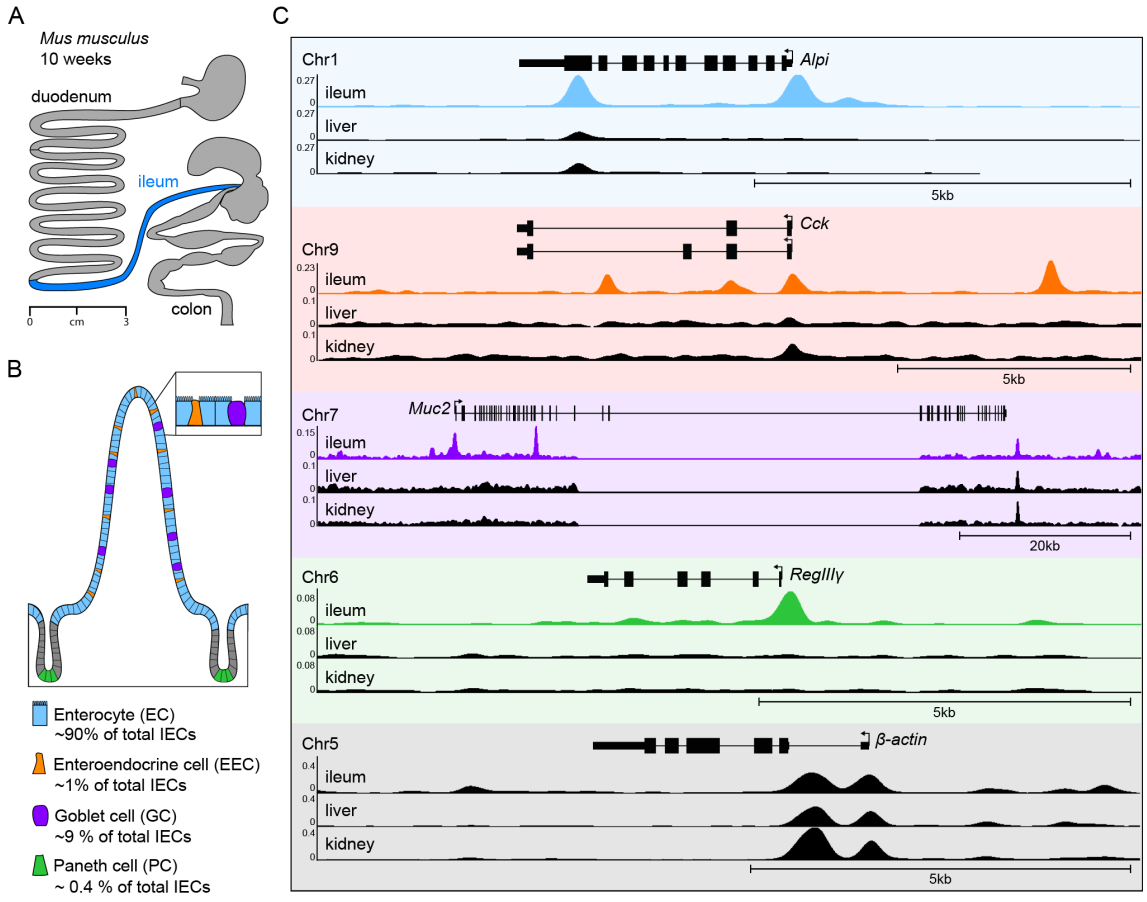


Figure 5.7: DNase-seq distinguishes cell-type specific open chromatin in the ileum
(A) Schematic of the mouse GI tract loosely drawn to scale and highlighting the ileum in blue. (B) A cartoon cross-section shows four differentiated cell types: absorptive enterocytes (light blue), enteroendocrine cells (orange), goblet cells (purple), and paneth cells (green) and approximate percentages of each cell type in the small intestine {Potten, 1998; personal communication with Rich von Furstenburg}. (C) Open chromatin at the *Intestinal alkaline phosphatase* (*Alpi,* enterocyte marker, light blue) locus, *Cholecystokinin* (*cck*, enteroendocrine cell marker, orange) locus, *Mucin 2* (*Muc2*, goblet cell marker, purple*)* locus, *Regenerating islet-derived protein 3 gamma* (*RegIIIγ*, paneth cell marker, green) locus, and *β-actin* (ubiquitously expressed, black) in the ileum, liver, and kidney. Note common and distinct peaks. There are no distinct tissue-specific peaks near the ubiquitously expressed gene *β-actin.*

kidney datasets (Figure 5.7C). Together, these data reveal that the ileum DNase-seq dataset from whole mIECs preparations is sensitive enough to detect putative cell type-specific and tissue-specific *cis*-regulatory modules for a variety of cell types within the intestinal epithelium at specific loci.

5.3.8 DNase-seq predicts transcription factors regulating intestinal gene expression

I next set out to explore the utility of the mIEC ileum DNase-seq dataset to discover transcription factors that regulate gene expression in the intestine. Out of the top 50,000 ileum peaks, there are 17,430 DH sites that have no overlap with the top 50,000 DH sites from liver or kidney (Figure 5.8A). To elucidate transcription factors that can function through the ileum-specific putative CRMs, I used MEME-Chip [309] to discover motifs *de novo* that are enriched in the top 50,000 ileum or 17,430 ileum-specific sequences and TOMTOM [243] to determine the similarity to a database of known transcription factor binding site motifs (Table 5.1). This analysis resulted in enrichment of a motif (iMEME 1) in the top 50,000 ileum DH sites that significantly matches binding sites for various nuclear receptors such as Hepatic nuclear factor 4-α, Peroxisome-proliferator activated receptors, and Farnesoid X receptor (FXR). Notably, a motif (iMEME 2) that is recognized by the gut enriched Krüppel-like factor 4 (Klf-4), as well a DREME motif that matches the binding site of GATA binding factors (Figure 5.8B) is enriched in the ileum-specific set of DH sites. One of the intestine-specific DH sites is located ~6.5kb upstream of the mouse *Angptl4* locus and the underlying sequence harbors a strong match to both the iMEME 1 and the DREME GATA motif (Figure 5.8C). Interestingly, similar TFBS motifs are required for intestine-specific reporter expression driven by the zebrafish in3.4 intronic intestinal enhancer presented in chapters 3 and 4 (Figure 3.6 and Figure 4.1)[290]. A recent ChIP-seq survey [145] of FXR binding in the ileum and liver revealed a strong FXR binding peak exactly overlapping the upstream intestine-specific DHS peak at the *Angptl4* locus (Figure 5.8C). Strikingly, neither the FXR binding peak nor the DHS peak were present in the respective liver datasets. Together, this data shows that motif prediction using the ileum DHS dataset can lead to

the discovery of *cis*-regulatory modules and associated transcription factors that potentially regulate intestine-specific gene expression.

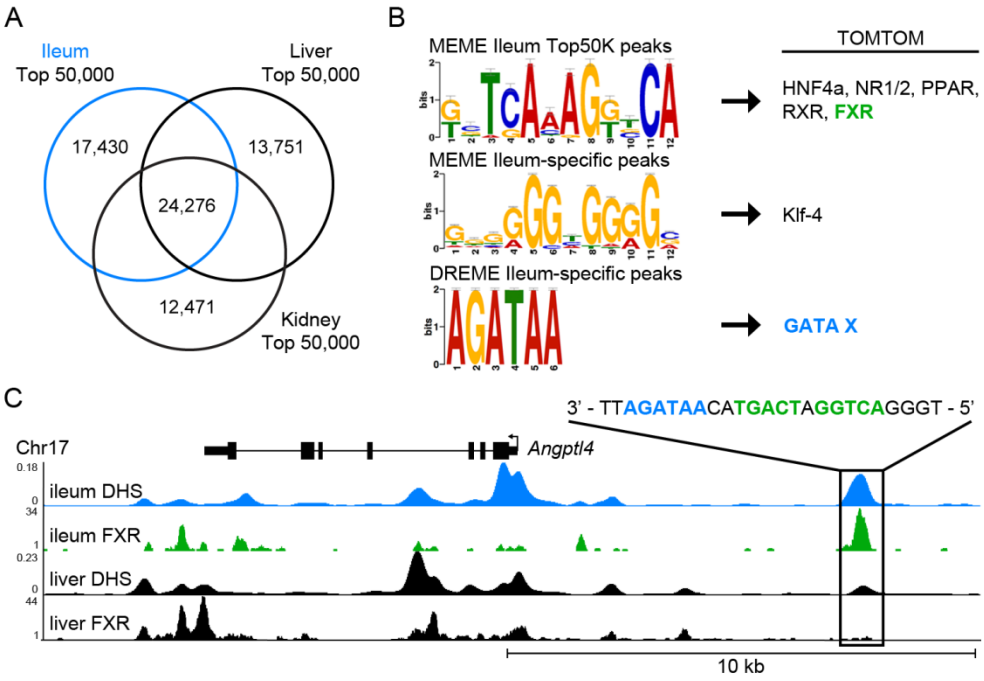


Figure 5.8: Ileum DNase-seq predicts motifs regulating intestinal gene expression
(A) Venn diagram showing the intersection of the top 50,000 DH peaks from ileum, liver, and kidney DNase-seq datasets. (B) *De novo* motif prediction (MEME-ChIP) using top 50,000 and the 17,430 ileum-specific DH sites uncovers motifs matching transcription factor (TF) binding sites of TFs that are known to regulate intestine gene expression. (C) An intestine-specific peak located upstream of the *Angptl4* TSS harbors a putative GATA binding site (blue) and a neighboring binding site that matches various nuclear receptor recognition sequences (green). A Farnesoid X Receptor (FXR) ChIP-seq peak [145] from the ileum, but not the liver, overlaps with the intestine-specific peak suggesting FXR may regulate intestinal expression via this upstream *cis*-regulatory module (CRM).

MEME motif	TOMTOM Hit	DREME Motif	TOMTOM Hit
MEME 1	NFE2L2, AP1, GCN4, ARG81, Fos	DREME 1	Klf4, SP1
MEME 2	Klf4	DREME 2	Abd-B, cad
MEME 3	Stat3	DREME 3	NFE2L2, AP1, GCN4, ARG81, Fos
MEME 4	AZF1	DREME 4	FEV, ELF5, SPI1, Stat3
MEME 5	?	DREME 5	HNF4A, NR2F1, PPARG:RXRA
MEME 6	?	DREME 6	Gata1, Tal1:Gata1
		DREME 7	E5, btn, zen, Scr, ftz, Antp, pb, lbl, eve, ind
		DREME 8	SP1
		DREME 9	NR4A2
		DREME 10	Myf, NHLH1
		DREME 11	PPARG:RXRA
		DREME 12	CREB1
		DREME 13	FOXA1, fkh, FKH2, FOXD1
		DREME 14	Gata1, Evi1, Tal1:Gata1
		DREME 15	MET28

Table 5.1: Motif prediction using 17,441 ileum-specific DH sites.
MEME-ChIP was used to search for *de novo* motif enrichment in ileum-specific DH sites (those sites that had no intersection with DH sites from liver and kidney). MEME-ChIP software performs motif analysis using two algorithms that find either long motifs (MEME) or short motifs (DREME). TOMTOM was used to determine if *de novo* motifs (MEME and DREME) were similar to transcription factor binding sites deposited in TRANSFAC and JASPAR databases.

5.3.9 DNase-seq predicts transcription factors regulating microbial response

I next used the DH sites near genes that are either up-regulated or down-regulated in the ileum of CONVD mice compared to GF controls [87] to predict transcription factors that regulate response to the microbiota. I first used the software Genomic Regions Enrichment Annotations Tool (GREAT) to analyze the functional categorization of hypersensitive sites near up- or down- regulated genes. Briefly, GREAT associates input genomic regions (for example DH or FAIRE sites) with genes by defining a “regulatory domain for each gene. This regulatory domain extends 5kb upstream and 1kb downstream from the TSS with extension up to the nearest next gene within 1Mb of the regulatory domain. GREAT uses these regulatory domains for each gene as the expected fraction of the genome that is associated with a given functional gene ontology annotation. GREAT then uses a binomial distribution to test for enrichment of input genomic regions over expected for each annotation category in

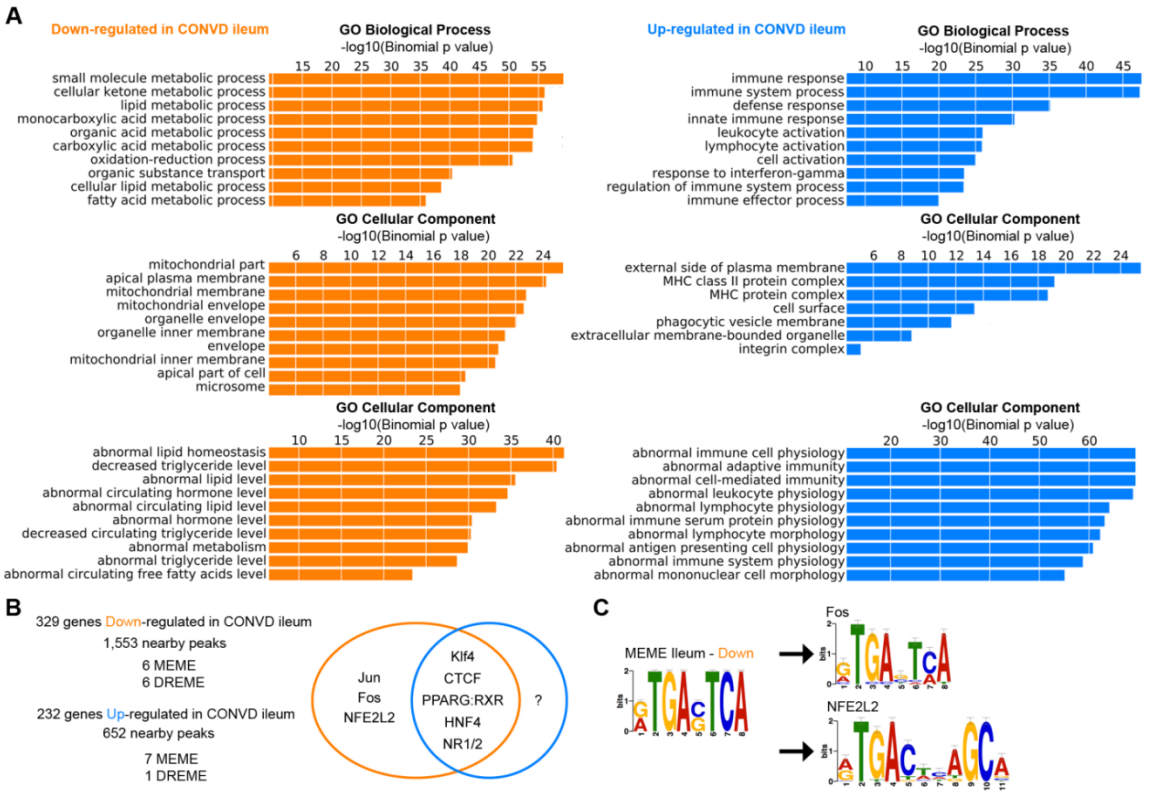


Figure 5.9: Motif prediction using DH sites near microbiota regulated genes
(A) Peaks near genes that are down- (orange) or up- (blue) regulated in the CONVD ileum were used for motif prediction. Down-regulated gene sets are enriched in functional annotation categories involved in cellular and nutrient metabolism, whereas genes that are up-regulated are involved in immune response. (B) *De novo* motif prediction (MEME-ChIP) uncovered multiple motifs that are common to both gene sets. A motif matching Gcn4, Jun, Fos, and NFE2L2 was enriched in peaks near genes down-regulated in the CONVD ileum.

order to functionally interpret the set of input genomic regions. GREAT analysis of DH sites near up- and down-regulated genes, though circular, reinforced the predictive capacity of GREAT and illustrated the different gene functions present in each category. For example, the genes that are down-regulated in the ileum of CONVD mice are enriched in GO categories and mouse phenotypes associated with lipid metabolic and other biosynthetic processes, whereas genes up-regulated in CONVD animals are associated with immune response (Figure 5.9A). I next used MEME-Chip/TOMTOM to search for over-represented motifs in each set. This analysis discovered a number of motifs that are present in both sets of peaks, despite the dramatically distinct set of

functions associated with each peak (Figure 5.9B and Table 5.2). Most motifs were similar to those discovered when using either the Top 50,000 ileum or the Ileum-specific DH sites (Table 5.1). However, this analysis uniquely identified MEME and DREME motifs that matched binding sites for Nuclear factor erythroid-derived 2 (NFE2L2 or Nrf2), a basic leucine zipper transcription factor known for its role in mediating gene response programs to oxidative stress. This is consistent with reports that colonocytes from GF mice are energy-deprived and have altered oxidative metabolic states [310]. Taken together, subcategorizing ileum DNase peaks using gene expression or non-overlap with other tissues has strong predictive capacity for a range of intestinal physiologies.

Down-regulated in CONVD ileum (329 genes, 1,553 peaks):

MEME motif	TOMTOM Hit	DREME Motif	TOMTOM Hit
MEME 1	AZF1	DREME 1	Klf4, SP1, MIG2, MIG3
MEME 2	Klf4, Trl1	DREME 2	FEV
MEME 3	CTCF, Mycn, Sna	DREME 3	Abd-B, cad
MEME 4	SP1, Klf4, MIG2, MIG3	DREME 4	GCN4, Fos, AP1, NFE2L2, ARG81
MEME 5	HNF4A, NR2F1, PPARG:RXRA	DREME 5	?
MEME 6	abi4, Myf	DREME 6	?

Up-regulated in CONVD ileum (232 genes, 652 peaks):

MEME motif	TOMTOM Hit	DREME Motif	TOMTOM Hit
MEME 1	SP1, Klf4, RGM1	DREME 1	?
MEME 2	CTCF		
MEME 3	PPARG:RXRA, NR1H2:RXRA, HNF4A, NR2F1		
MEME 4	PPARG:RXRA, NR1H2:RXRA, HNF4A, NR2F1		
MEME 5	?		
MEME 6	?		
MEME 7	?		

Table 5.2: Summary of motif prediction using DH sites near microbiota regulated genes
MEME-ChIP was used to search for motif enrichment using DH sites (from the top 50,000 ileum set) that are near genes down-regulated or up-regulated in the CONVD ileum. TOMTOM was used to determine if *de novo* motifs (MEME and DREME) were similar to transcription factor binding sites deposited in TRANSFAC and JASPAR databases. Question marks indicate TOMTOM did not predict a TFBS that matched that MEME or DREME motif.

5.3.10 FAIRE-seq uncovers ancient CRMs in the zebrafish

Two of the major motivations for performing FAIRE-seq in the adult zebrafish intestinal epithelium were (i) to create a map of active *cis*-regulatory DNA to combine with primary sequence conservation information in order to guide *in vivo* structure/function assays such as those described in Chapter 3 and (ii) to understand the evolution of intestinal CRMs through comparisons with open chromatin in mouse IECs. We computed the overlap of adult zebrafish FAIRE-seq peaks (zIEC 1, zIEC2, the intersection, and the union) with zebrafish Conserved Non-genic Elements (zCNEs) [311] (Figure 5.10A and Table 5.3). Each zCNE was defined as being conserved to at least two species with at least 65% sequence identity to two other species in an alignment between 15 vertebrate species for at least 50 bp. The zCNE set consists of 54,533 elements, 12,778 of which are conserved with humans or mouse. Note that the relatively low number of zCNEs is due to the lack of genome sequences from fish species at appropriate evolutionary distances from the zebrafish to permit detection of sequence conservation. There is variability in the FAIRE datasets some of which is attributable to slightly varying the experimental conditions. However, the union of zIEC1 and zIEC2 captured the most zCNEs and I will discuss these results. Of 124,466 FAIRE-seq peaks, only 5,931 (4.8% of total FAIRE peaks, 11% of total zCNEs) overlapped with the zCNE set and 1,546 (1.2% of total FAIRE peaks, 12% of total Hs or Mm conserved zCNEs) overlapped with zCNEs conserved with human or mouse (Figure 5.10A). The results were mostly similar for the intersection of zIEC datasets (those regions that were discovered in both FAIRE-seq experiments) or either zIEC FAIRE-seq dataset alone (Table 5.3). In any case, conservation could not predict 95% of FAIRE-seq peaks. Assuming that zIEC FAIRE-seq captures the functional regulatory landscape similar to published reports, this data corroborates results presented in Chapter 3 questioning the

utility of conservation with currently sequenced genomes for predicting non-developmental CRMs in the zebrafish.

Taken from a different perspective, this analysis yielded 1,443 FAIRE-seq non-genic peaks that distinguish chromatin regions active in the intestinal epithelium that have been conserved for ~450 million years since the last common ancestor of fish and mammals. I used GREAT to associate these peaks to putative target genes and analyze the functional enrichment. This revealed enrichment near genes involved in transcriptional regulation and synthesis of macromolecules, as well as various signaling pathways known to function in development and intestinal epithelial biology such as nuclear receptors and Wnt signaling (Figure 5.10C). These FAIRE sites are enriched near genes that cause defects in the intestine, gut epithelium, or endocrine systems when mutated (Figure 5.10C). I cross-referenced regions in each functional annotation category with genes regulated by the microbiota in the mouse intestine [87,195] and involved in nutrient metabolism. Notably, *Elongation of long-chain fatty acids family member 6* (*elovl6*) is an enzyme involved in the elongation of saturated and monounsaturated fatty acids and has been implicated in diseases such as diabetes and obesity [312]. FAIRE-seq revealed a region of open chromatin within the first intron of zebrafish *elovl6* that directly overlapped with the highly conserved zCNE. Inspection of the *Elovl6* locus in mouse also revealed a strong DNase hypersensitive peak that overlapped the conserved region (Figure 5.10D). *Hepatic nuclear factor 4 alpha* (*Hnf4*α) is a well-studied transcription factor involved in the development and maintenance of the liver and intestine. There is a highly conserved region upstream of the *Hnf4*α TSS that overlaps open chromatin in both the zebrafish intestinal epithelium and the mouse intestinal epithelium (Figure 5.10E). These data highlight only 2 of over 1,000 regions conserved between mouse and fish that are putative intestinal *cis*-regulatory modules in

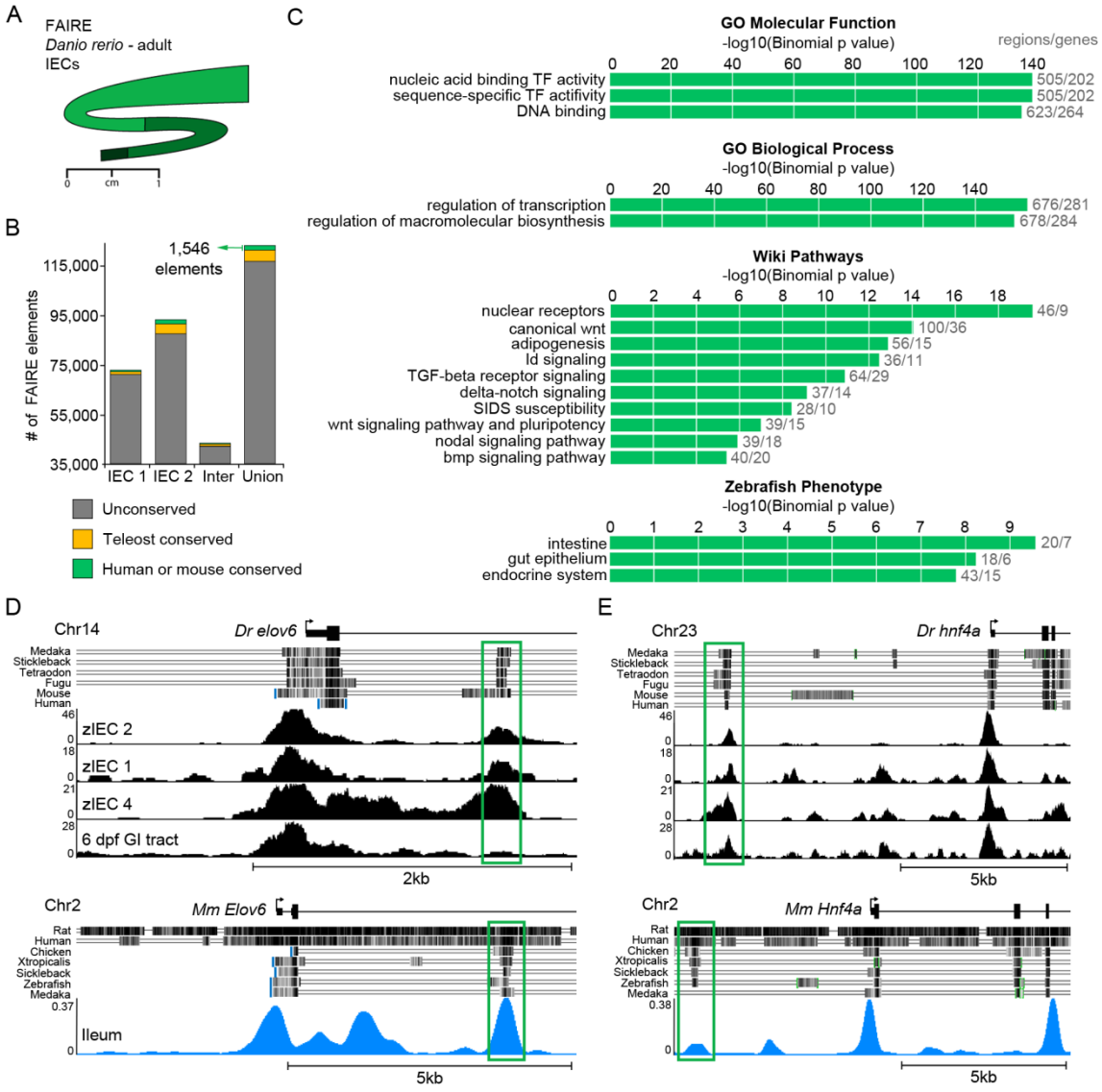


Figure 5.10: FAIRE-seq in zebrafish IECs uncovers ancient CRMs
(A) Cartoon schematic of the zebrafish intestine used for zIEC FAIRE-seq drawn to approximate scale. Anterior intestine (light green), middle intestine (green), and posterior intestine (dark green). (B) The overlap of zebrafish conserved non-genic elements (zCNEs) [311] and FAIRE-seq peaks from zIEC 1, zIEC2, the intersection of zIEC 1/zIEC 2, and the union. Gray represents non-overlap or unconserved FAIRE peaks. Orange represents FAIRE peaks conserved with teleosts and green represents FAIRE peaks conserved with mouse (*Mm*) or human (*Hs*). See Table 5.2 for more detail. (C) FAIRE-seq peaks conserved with human or mouse were linked to nearest genes and analyzed for functional enrichment using GREAT. *Hs*/*Mm* conserved peaks were enriched near genes involved transcriptional regulation and developmental signaling processes. (D) UCSC browser view of conservation and IEC open chromatin tracks at the *elongation of fatty acids 6* (*elov6*) locus from zebrafish intestine (FAIRE-seq) and mouse ileum (DNase-seq) reveals a highly conserved peak in the 1st intron. (E) UCSC browser view of conservation and IEC open chromatin tracks at the *hepatic nuclear factor 4* α (*Hnf4*α) locus from zebrafish intestine (FAIRE-seq) and mouse ileum (DNase-seq) reveals a highly conserved peak in the upstream of the TSS.

both lineages and can now be used as a guide for in depth structure-function analysis in the zebrafish. The extent of divergent, convergent, and parallel evolution of tissue-specific and microbial control of transcriptional regulation in the intestine will soon be open to in depth interrogation.

Datasets	# elements	Bases of zebrafish genome covered (bp)	% of zebrafish genome	FAIRE peaks and zCNE overlap (elements)	FAIRE peaks and zCNE overlap (bp)	FAIRE peaks and zCNE overlap (% of zebrafish genome)	FAIRE peaks and zCNE overlap (% of FAIRE peaks)	FAIRE peaks and Hs/Mm-zCNE overlap (elements)	FAIRE peaks and Hs/Mm-zCNE overlap (bp)	FAIRE peaks and Hs/Mm-zCNE (% of zebrafish genome)	FAIRE peaks and Hs/Mm-zCNE (% of FAIRE peaks)
zCNE	54,533	6,643,241	0.470								
Hs/Mm-zCNE	12,778	2,256,568	0.160								
IEC 1 peaks	73,570	28,142,189	1.992	1,793	177,965	0.013	2.437	438	53,785	0.004	0.595
IEC 2 peaks	94,652	37,216,904	2.635	5,560	643,107	0.046	5.874	1,443	208,080	0.015	1.525
Intersection (IEC 1:IEC 2)	43,749	15,324,236	1.085	1,418	142,981	0.010	3.241	334	42,165	0.003	0.763
Union (IEC 1:IEC2)	124,466	50,034,857	3.542	5,931	678,091	0.048	4.765	1,546	219,700	0.016	1.242

Table 5.3: Summary of the intersection of FAIRE-seq peaks with zebrafish conserved non-genic elements (zCNEs)
The intersection (overlap) of FAIRE sites from two replicates of zebrafish intestinal epithelial cells (IEC1 peaks and IEC2 peaks) with zCNEs is shown. The intersection and union of FAIRE IEC 1 and IEC 2 datasets was also determined and the intersection with zCNEs computed from these sets. The data sets are represented as the number of elements, bases, and the percentage of the zebrafish genome covered by each set. The intersection of the sets is represented by the number of elements, bases, the percentage of the zebrafish genome, and the percentage of total FAIRE peaks covered by each intersection. Two categories are shown: the overlap of FAIRE peaks with all zCNEs and the ovlerlap of zCNEs that are conserved with human (Hs) or mouse (Mm).

5.5 Discussion

5.5.1 Genomic atlas of open chromatin in the vertebrate intestinal epithelium

Non-genic *cis*-regulatory DNA modules govern cell type-specific identity and response to environmental factors such as the intestinal microbiota. Predicting CRMs using only sequence conservation largely ignores context. Here I have applied DNase-seq or FAIRE-seq to the mouse and zebrafish in order to elucidate the open chromatin landscape in IECs under varying environmental conditions (Figure 5.3-5.5). I presented pilot results from conventionally-raised mouse and zebrafish IECs that exhibited hallmark features of published datasets and corroborated known intestine-specific

regulatory regions at the mouse *Villin1* locus and the zebrafish *angptl4* locus. I also showed that these datasets could be used to predict novel CRMs at numerous other loci (Figure 5.7, 5.10). Although my experimental approach was to limit the cellular complexity of the sample through epithelial sloughing and microdissection, the heterogeneity of cell types covered in these datasets presents both an opportunity and a challenge. It is expected that each intestinal epithelial cell type should have a somewhat unique regulatory landscape and therefore a distinct potential to mediate microbiota-associated responses. I chose to sacrifice complete homogeneity for breadth and included all cell types within the intestinal epithelium as opposed to sorting specific cell types. In this way there is increased coverage of potential microbial responses while limiting the time cells would spend outside of their native environment prior to lysis or fixation. In the future, it would be interesting to parse out crypt versus villus regulatory regions using gene expression profiles from intestinal crypts or villi [313]. There are also robust expression datasets available comparing paneth cells and Lgr5+ stem cells [314]. I am currently unaware of any available gene expression datasets of sorted goblet cells, enteroendocrine cells, or enterocytes though these may exist and would be very helpful in assigning peaks to cell type-specific gene expression. Finally, we should soon have datasets from the mouse colon and duodenum and it will be interesting to compare the regulatory landscape in three similar cell populations distributed along functionally distinct regions of the alimentary tract. Naturally, biological replicates will enable corroboration of tissue-specific putative CRMs highlighted in this Chapter.

5.5.2 Open chromatin maps to predict transcription factors

I showed that sequences distinguished by DNase hypersensitivity could predict transcription factors that have known roles in IEC biology and likely function through these DNA regions to regulate gene expression in the ileum (Figure 5.8, 5.9). There are

a number of important considerations for future efforts aimed at motif prediction using these datasets. All points are equally valid for zebrafish FAIRE-seq data, though I focus on the DNase-seq mIEC data for brevity.

The first point pertains to the specific sequences used for motif predictions. The peak calling method is very important and is an active area of research. In this study, we used the statistically rigorous F-seq package [301] to generate discrete peaks by fitting sequencing tag-based signal data to a gamma distribution and determining the signal value of each distribution. DNase hypersensitive site status can then be designated by a certain p-value cut-off or as a certain quantile of total discrete peaks. In general, a p-value less than 0.05 or the top ~100,000 peaks have been used arbitrarily as cutoffs for using peaks in *in silico* experimental analyses [141]. However, the selected cut-off can likely have strong impact on the analysis. For example, using the top 25,000 peaks would enrich for promoter associated DH sites (Figure 5.6A). Furthermore the algorithm used for determining the optimal peak boundary, or bandwidth, will affect the results. On a first approximation, F-seq peak calls were narrower and distinguished neighboring DNase hypersensitive peaks better than MACS, an alternate method of peak calling designed using ChIP-seq datasets (data not shown) [315]. However, in browsing peak calls in the DNase-seq dataset, I noticed multiple occurrences in which neighboring CRMs could be visually distinguished as two distinct peaks yet both F-seq and MACS combined these peaks into a single DNase hypersensitive site. Also, there were occasions (such as *Sox2* and *Sox9*, not shown) in which the peak call covered most of the gene, though an easily distinguishable peak at the promoter was observed. One way around the bandwidth problem would be to use only a set number of bases flanking the peak summit or peak center as the peak boundary. This would also standardize the number of bases used per sequence in downstream *in silico* analyses. However, similar

to the ‘Goldilocks conundrum’, too broad will include non-functional DNA and too narrow will exclude functional DNA.

The second major point involves using gene sets to filter peaks and associated sequences used for motif analysis. It is non-trivial to determine the gene regulated by a given CRM as distinguished solely by DNase hypersensitivity. Current methods (such as those used in this study) most often assign peaks to the nearest one or two genes as this is an easy solution and requires no additional information other than an annotated genome. However, this underestimates the complexities of gene regulation as enhancers or other *cis*-regulatory features can frequently be distal from target genes [68]. Nearest gene methods ignore recent observations that the three-dimensional organization of genomes have a profound impact on gene regulation through long-range promoter-promoter and promoter-enhancer interactions [316]. Furthermore, a recent comprehensive analysis of *cis*-regulatory sequences in multiple tissues from the mouse showed strong evidence that the genome is partitioned into functional domains in which CRMs are coordinately regulated [317]. Despite these considerations, there is ample data that suggest CRMs located near genes often regulate those genes and can be used for generating testable hypotheses. Indeed using basic approaches, I was able to identify a number of candidate factors that may regulate microbial control of gene expression through DH sites in the ileum (Figure 5.9). Including RNA expression data or incorporating published chromatin interaction datasets from the mouse intestine [317] may help mitigate the underlying assumptions of nearest gene approaches and sidestep the need to generate chromatin interaction datasets from IECs.

Finally, it is important to highlight that the particular method or software used for motif enrichment analysis is subject to debate. There is what seems to be an endless stream of motif prediction software available, each with apparent positive and negative features. In this study, I used MEME-ChIP because it allows *de novo* motif generation

from ~100,000 input sequences, integrates comparison against databases of known motifs, and utilizes a graphical user interface. However, an important drawback is that MEME-ChIP does not sample every sequence in the input query, but instead randomly samples the center 100 bp from up to 600 sequences. Alternatively, DREME (in the same package) samples the center 100 bp from all sequences, but is biased towards short motifs. MEME-ChIP does allow inclusion of a background Markov model generated from random sequences, though I did not utilize this feature in my preliminary analysis. There are multiple other packages such as Homer [318], cisFinder [319], CENT-DIST [320] and others [321] that should be explored. Importantly, incorporating genome-wide DNase-footprinting analysis [142] into the motif search will utilize the full capacity of the DNase-seq datasets. However, all efforts toward motif prediction are subject to the same biases discussed in Chapter 4, and therefore dependent on both the quality of the input sequences as well as the queried databases.

5.5.3 Integrating zebrafish and mouse open chromatin maps

Approximately 450 million years distinguish the independent evolutionary histories of mice and zebrafish [204]. Most comparisons between zebrafish and mammals have focused on CRMs regulating developmental processes as many of these regions are highly conserved and therefore discernable through comparative sequence analysis [124]. To my knowledge the datasets generated in this study mark the first time that the chromatin landscape of a common differentiated tissue from such distant vertebrate relatives has been profiled and compared. I showed that the FAIRE-seq zIEC dataset could be used to elucidate ancient non-genic DNA likely functional in both the mouse and zebrafish intestine (Figure 5.10). However, conserved FAIRE-seq peaks are the vast minority. It has been shown that there are strong evolutionary constraints that maintain tissue-specific gene expression patterns across vertebrates despite an

apparent absence of sequence conservation in non-genic DNA to account for conserved gene expression [322,323]. Comparing open chromatin landscapes at conserved genes with high expression in the intestine, but with no apparent non-genic sequence conservation, should present valuable opportunities for understanding the evolution of vertebrate *cis*/*trans* regulatory systems.

How does the location of CRMs relative to the gene body change over evolutionary time? Does TFBS composition within functional modules turnover or change altogether? Does gene function or cell type correlate with differential conservation of CRMs active in the intestine? Recently there has been some progress in addressing these questions in vertebrates [120,324], however most of the analysis has been limited to computation and lack appropriate heterologous reporter systems for cross-species comparative analysis. In this light, the *Angptl4* gene provides an illustrative example of how mouse and zebrafish data can be integrated. I showed in Chapter 3 that an unconserved zebrafish intronic CRM recapitulates intestine-specific and microbial suppression of zebrafish *angptl4* transcription. The orthologous intron did not drive intestinal expression in the zebrafish, despite strong mammalian conservation in the same intron and known regulatory potential (Figure S3.8). DNase-seq in mIECs allowed the discovery of a weakly conserved DH site upstream of the mouse *Angptl4* TSS that is targeted by the transcription factor FXR (Figure 5.8) and has a binding site composition similar to the zebrafish in3.4 module. It will be interesting to determine if this putative mouse CRM is sufficient to drive expression in the zebrafish and/or mouse intestinal epithelial cells. It is possible that despite numerous reports of conserved regulatory function in the absence of sequence conservation [196,323], cross-species functional conservation will be restricted in this case. In any result, this will be an informative case study toward understanding the evolution of transcriptional regulation of

non-developmental genes with conserved tissue-specific and environmental control expression patterns.

Though tedious, assaying conserved and un-conserved candidate CRMs *in vivo* is a powerful avenue to uncover the intricacies of intestine-specific regulatory logic. I estimate that using current techniques and vector systems, one person could clone and assay 50 putative CRMs in the zebrafish over the course of a couple of months. The current Tol2 vector system [186] is the major limiting factor, as it requires a two-step cloning procedure. The throughput could be dramatically increased if alternative single-step cloning systems [325] or perhaps non-vector systems [124,326] were explored. Also, it would be highly beneficial to have a comparable *in vivo* assay system available in the mouse model. Transgenic reporter expression via pronuclear injections is the gold standard for assaying CRMs in the mouse [123], but it is time consuming and expensive. An alternative method could be explored in which constructs are electroporated into the embryonic mouse intestine [327] or transfected into organoid culture systems [328]. Caco-2 or other cell culture lines are less desirable though adequate alternatives for CRM validation.

Current methods to directly compare cross-species chromatin genomic datasets to identify shared and distinct peaks genome-wide have required sequences to be aligned to a single reference genome [329] or to convert genome coordinates of peak calls to genome coordinates of another organisms using sequence alignment. The evolutionary distance between mouse and zebrafish may make these tasks difficult. In any case, I was able to use the overlap of FAIRE-seq peak calls with annotated sets of zebrafish conserved non-genic elements to discover putative functional CRMs conserved between mouse and zebrafish (Figure 5.10). Computing the reciprocal set (sequences distinguished by mouse DH sites that are conserved with zebrafish) and finding the overlap with zIEC FAIRE-seq peaks will elucidate open chromatin sites

conserved in both the mouse and fish IECs. However, this will ignore open chromatin regions that are not distinguished by primary sequence similarity. Alternative approaches using transcription factor binding site composition and clustering may greatly extend the set of functional CRMs conserved between zebrafish and mouse [266,330].

5.5.4 Microbial impact on *cis*-regulatory function and evolution

The primary motivation for this project was to understand how the microbiota has impacted the functional landscape of non-genic DNA in the intestinal epithelium. One of the most interesting questions that these datasets should address is how a cell uses the genome to respond to dramatically different environmental conditions. In addition to the CONV-R mouse and zebrafish samples analyzed here, we have multiple replicates of CONV-R and GF samples in the sequencing pipeline that are expected to provide answers in the coming months. The degree of effect of microbial colonization on chromatin openness and CRM activity remains unknown. It is possible, though doubtful, that we will observe large or binary changes in open chromatin landscapes in GF vs. CONV-R environmental conditions. It is more likely that differences will be moderate and it is expected that the high signal to noise ratio and quantitative nature of the DNase I hypersensitivity assay will be particularly useful for discerning regions that mediate response to microbial activity. Furthermore, it may be that differential DNase footprinting [142] rather than overall chromatin accessibility can explain gene expression differences in GF vs. CONV-R animals. There is precedence to suggest that differential DNase digestion can explain many of the gene expression changes observed between species, cell-types, and environmental conditions [331,332]. Using rigorous statistical methods to discern regions of the genome that are more or less hypersensitive in GF compared to CONV-R animals and correlating these changes with gene expression will be paramount to the success of this project. Comparing microbiota regulated chromatin in IECs to other

tissues such as liver or kidney will help parse out the overlap between tissue-specificity and environmental responses.

One of the most exciting questions that this study should help address is where in evolution did the regulatory regions that mediate host responses to the microbiota emerge and how fast do they turnover? To answer this question we will need to define a set of putative CRMs as characterized by differential DNase hypersensitivity or FAIRE enrichment and use sequence conservation thresholds in these regions to determine at what branch in the phylogeny each CRM appeared [120]. Again the mouse DNase-seq data will be particularly informative because of the multitude of mammalian genomes sequenced. Further binning CRMs into functional categories using GREAT, filtering for tissue-specificity, and defining the branch of innovation should elucidate salient features of *cis*-regulatory evolution. Combining zIEC and mIEC genomic datasets with (i) the utility of the zebrafish as a gnotobiotic system amenable to rapid genetic manipulation and (ii) the biomedical relevance and extensive reagent resource of the mouse model, is expected to provide unprecedented insight into the evolution of host-microbe symbiosis.

Sample	Total Raw Reads	Read Length (bp)	Total Aligned Reads	Unique Alignments	Usable alignments	Percent of Raw Reads Successfully Aligned
Ileum mIEC DNase-seq CONV-R Rep1	199,593,023	20	156,832,519	142,085,508	143,249,947	91.3

Sample	Total Raw Reads	Read Length (bp)	Total Reads Passing TagDust	Median Read Quality Score	Total Aligned Reads	Percent of Raw Reads Successfully Aligned	Signal-to-Noise Score Quartile 1	Quartile 2	Quartile 3	Interquartile range	Mean
FAIRE-seq zIEC 1	51,228,513	50	51,131,614	37.84	24,022,096	46.89	8.67	11.00	16.00	7.33	15.75
FAIRE-seq zIEC 2	60,412,982	50	60,096,686	37.78	30,214,881	50.01	9.50	16.00	36.00	26.50	28.92
FAIRE-seq zIEC 3	100,983,905	50	99,790,941	37.49	42,194,968	41.78	10.20	12.33	16.00	5.80	19.21
FAIRE-seq zIEC 4	79,683,638	50	79,578,183	37.47	43,963,433	55.17	5.75	7.25	9.50	3.75	9.76
6 dpf GI tracts CONV-R	46,976,867	50	46,918,010	37.69	24,158,191	51.43	6.50	8.00	10.00	3.50	10.68
6 dpf GI tracts GF	32,089,154	50	32,046,487	37.69	16,896,429	52.65	6.00	7.00	8.00	2.00	8.96

Table 5.4: DNase-seq and FAIRE-seq sequencing results summary
Illumina sequencing reads and alignment statistics for CONV-R mouse ileum DNase-seq replicate 1, CONV-R adult zebrafish IECs FAIRE-seq replicates 1-4, CONV-R 6dpf zebrafish GI tract FAIRE-seq replicate 1, and GF 6dpf zebrafish GI tract FAIRE-seq replicate 1 are shown.

5.6 Materials and Methods

5.6.1 Mouse and zebrafish husbandry

All mice used in this study were in the C57BL/6 background sourced from Jackson Laboratories and reared under germ-free or specific pathogen-free (SPF) conditions in the National Gnotobiotic Rodent Resource Center at the University of North Carolina (UNC) at Chapel Hill. Production, colonization, maintenance, feeding, and sterility testing of germ-free mice were performed using standard procedures per the National Gnotobiotic Rodent Resource Center. Animals were housed on Alpha-dri bedding (Shepherd) and fed 3500 Autoclaveable Breeder Chow (Prolab) ad libitum.

All zebrafish used in this study were wild type TL strains reared in the Zebrafish Aquaculture Core Facility at UNC Chapel Hill. Conventionally raised adult fish were fed twice daily with Great Salt Lake strain brine shrimp (*Artemia*, Aquafauna Bio-Marine, ABM-GSL-TIN-90) supplemented with flake food (5 parts Tetramin ®Flakes Aquatic Ecosystems, 16623; 1.5 parts Zeigler ® Aquatox Flakes, Aquatic Ecosystems, AX5; 1.5 parts Spirulina Flakes Aquatic Ecosystems, ZSF5; 1 part Cyclop-eeze Argent Chemical Laboratories, F-CYCL-FD30-CS; 1 part San Francisco Bay Freeze-dried brine shrimp (Aquatic Ecosystems, SB113. Conventionally raised 6 dpf zebrafish were offspring from TL strains reared and maintained as described [226]. Production, colonization, maintenance, and sterility testing of 6 dpf germ-free zebrafish were performed as described [57,94].

All experiments using mice and zebrafish were performed according to established protocols approved by the Animal Studies Committee at UNC at Chapel Hill.

5.6.2 DNase hypersensitivity

IECs isolation from mouse

Eight-week old mice were terminally anesthetized with 0.5 ml isofluorane in an airtight container and euthanized via cervical dislocation. Duodenum (anterior 5 centimeters of midgut), ileum (posterior 6 centimeters of midgut), and colon (6 centimeters of terminal hindgut) were harvested and placed into three separately labeled 50ml conical tubes containing ice-cold PBS. Using a dissecting scope, intestine-associated mesentery, adipocytes, and blood vessels were removed from each segment. Using scissors and starting with colon, each segment was splayed, vigorously washed quickly five to ten times with 50 ml ice-cold PBS, and transferred into dissociation reagent 1 (DR1; 30 mM EDTA, 1.5 mM DTT, 0.5x Complete protease inhibitors (Roche), in 1x PBS) for 15 minutes on ice. Segments were transferred to Dissociation Reagent 2 (DR2; 30 mM EDTA, 0.5x Complete protease inhibitors (Roche), in PBS) and moderately shaken by hand for 5-20 minutes (depending on tissue) until most epithelial cells are sloughed [291]. Note that isolation of intestinal epithelial cells were staggered at 5 minute intervals because the colonic epithelium takes approximately 15 minutes to slough, the ileal epithelium takes 10-15 minutes, and the duodenal epithelium takes 5-10 minutes. Intestinal lamina propria was removed and 8 ml of cold PBS was added to the cells on ice. Care was taken to minimize contact of cells with polypropylene pipette tips and polystyrene pipettes due to cell stickiness and loss. Cells were pelleted at 400 x G at 4°C and washed twice with 13 ml cold PBS. During each wash cells were resuspended by flicking in 1 ml PBS before adding the remaining 12 ml. After the 2nd wash, cells were resuspended in 0.5 ml cold PBS and 0.1 ml was reserved for RNA extraction.

DNase hypersensitivity assay on mouse IECs

Cells were gently lysed by adding 10 ml 0.1% Igepal in Resuspension Buffer (10 mM Tris-Cl, pH 7.4, 10 mM NaCl, 3 mM $MgCl_2$) containing 1x Complete Protease Inhibitors, inverted three times, and three gentle shakes. Nuclei were pelleted at 600 x G for 10 min at 4°C. Nuclei were resuspended by flicking in 0.73 ml RSB. Nuclei aliquots (0.12 ml) were transferred to labeled 1.5 ml eppendorf tubes on ice using a wide bore pipette tip (cut with razor). Nuclei were incubated at 37°C for 30 seconds, 1 minute, 2 minutes, 4 minutes, or 8 minutes and reactions stopped by addition of 0.33 ml cold 50mM EDTA. Agarose plugs were made by pipetting 0.45 ml of 55°C 1% low-melting point agarose (in sterile 50 mM EDTA, pH 8.0; InCert, Lonza, 50121) directly to the reactions on ice and quickly distributing approximately 80 µL per plug yielding about 10 plugs per time point. Plugs were solidified at 4°C and transferred to 50 ml of LIDS Buffer (10 mM Tris-Cl, pH 8.0; 1% lauryl sulfate lithium salt (Sigma); 100 mM EDTA) and shaken for 1 hour at room temperature at 60 rpm. LIDS Buffer was changed and plugs were incubated overnight at 37°C. Plugs were washed five times with 50 ml of 50 mM EDTA for 1 hour each wash. Plugs were then stored at 4°C in 50 mM EDTA. Half of one plug at each condition was used to determine the appropriate amount of digestion to be used for constructing sequencing libraries. Libraries were made exactly as described [137].

5.6.3 Formaldehyde Assisted Isolation of Regulatory Elements (FAIRE)

Isolation of IECs from adult zebrafish

Approximately 1-year post fertilization female zebrafish (TL strain) were terminally anesthetized with 0.8% tricaine (w/v) (Argent Chemical Laboratories). The midgut or midgut/hindgut segments were dissected, splayed, and washed extensively

with ice-cold 1x PBS with care taken to remove as much intestine associated fascia, adipocytes, and blood vessels as possible. Three washed intestines were transferred into dissociation reagent 1 (DR1; 30 mM EDTA, 1.5 mM DTT, 0.5x Complete protease inhibitors (Roche), in 1x PBS) for 15 minutes on ice. Segments were transferred to Dissociation Reagent 2 (DR2; 30 mM EDTA, 0.5x Complete protease inhibitors (Roche), in PBS) and moderately shaken by hand for 5 minutes until most epithelial cells were sloughed [291]. Intestinal lamina propria was removed and 8 ml of cold 1x PBS was added to the cells on ice. Care was taken to minimize contact of cells with polypropylene pipette tips and polystyrene pipettes because cells would stick to the sides. Cells were pelleted at 500 x G at 4°C and washed once with 13 ml of cold PBS. Cells were re-suspended by flicking in 0.5 ml cold 1x PBS and 0.1 ml was reserved for RNA extraction.

FAIRE on adult zebrafish IECs

FAIRE was performed exactly as described [139] with the following modifications. Briefly, freshly isolated intestinal epithelial cells were directly fixed for 5-10 minutes in 10 ml of 1-3% w/v Formaldehyde solution (in 1x PBS) at room temperature and gentle rocking. Glycine (2.5M) was added to a final concentration of 125 mM to quench the formaldehyde. Cells were pelleted at 600 x G and washed three times in cold 1x PBS without dissociating the pellet. Note that dissociation of the pellet often resulted in significant cell loss due to sticking to the sides of the conical tubes. Fixed and washed cell pellets were flash frozen and stored at -80°C. Cells were lysed in 2 ml Lysis Buffer A (10 mMTris-HCl (pH8.0), 2% (vol/vol) Triton X--100, 1% SDS, 100 mM NaCl and 1 mM EDTA) and sonicated using a Branson Sonifier 450D equipped with a microtip for 6-13 cycles (1 second burst, 0.5 second pause, for 30 seconds/cycle at 70% intensity)

allowing samples to cool on ice for 1 minute between cycles. The rest of the protocol was performed exactly as described [139].

FAIRE on dissected 6 dpf GI tracts.

For each condition, four rounds of twenty-five GI tracts were dissected from 6 dpf zebrafish larvae into 0.1 ml of 1x PBS on ice (30 minutes per 25 GI tracts) totaling approximately 100 GI tracts per FAIRE experiment. After each round, the GI tracts were fixed for 10 minutes in 4% Formaldehyde solution (in 50 mM HEPES, 100 mM NaCl, 1 mM EDTA), quenched with 125mM Glycine (final concentration) for 10 minutes at RT, washed three times in 1x PBS at 4°C, flash frozen and stored at -80°C. The GI tracts were then lysed in 2 ml of Lysis Buffer A and sonicated as described above. The rest of the protocol was performed exactly as described [139]. Quantitative PCR was performed as described above using primers listed in Table S3.2.

5.6.4 Library preparation and high throughput sequencing

DNase-seq libraries were prepared exactly as described [137] with 5 μg of DNA pooled from nuclei digested for 2 minutes, 4 minutes, and 8 minutes at 37°C for mouse and 0 minutes, 2 minutes, and 4 minutes at 25°C for zebrafish by Chris Frank in the lab of Greg Crawford. The ileum DNase-seq library was sequenced at the Duke Sequencing and Analysis Core Resource using Illumina HiSeq. Sequencing generated 156,832,519 mapable reads of read length 20 bp that were aligned to the mouse genome (MCBI37/mm9) using BWA [333] with default parameters. Raw aligned sequencing reads were smoothened using a kernel density estimation function called Parzen windowing [300,301], which allows identification of DNase hypersensitivity (DH) sites from uniquely mapped tags. DNase hypersensitive sites (or peaks) used for analysis

were called using F-seq [301]. Peaks were also called using MACS [315] as comparison but data was not shown using these peaks. Alignment and peak calling was performed by Yoichiro Shibata in the lab of Greg Crawford. A summary of the sequencing reads and mapped alignments are can be found in Table 5.4.

FAIRE-seq libraries were prepared using the TruSeq kit (Illumina, Cat. #15025064) according to manufacturer's specifications with the following exceptions. 100 ng of input FAIRE DNA was used for all zIEC samples and 60 ng were used for 6dpf GI tracts. Adaptors were diluted 1/10 prior to ligation. Libraries were verified using an Agilent Bioanalyzer by the UNC Bioinformatics and Genomics Core facility and sequenced (two libraries multiplexed per lane) using Illumina HiSeq at the UNC High-throughput Sequencing Core Facility. FAIRE-seq sequencing results were processed and mapped to the zebrafish genome (Zv9/GCA_000002035.2) using Zinba [334] by Jeremy Simon in the lab of Jason Leib. FAIRE-seq peaks were called using MACS [315]. A summary of the sequencing reads and mapped alignments are can be found in Table 5.4.

5.6.5 Bioinformatic analysis of DNase-seq and FAIRE-seq datasets

Most basic computational tasks were performed using the "operate on genomic intervals toolset" in Galaxy [335 2010,336] or the "integrative analysis" toolset on Cistrome [337]. Distributions of DHSs across genomic features (promoter, exons, introns, etc.) were computed with help from Chris Frank using Refseq gene annotations and an in-house script in the Crawford lab. Relative gene expression scores from wildtype C57BL/6 CONV-R ileum [87] were sorted and correlated with open chromatin at the TSS using R with the assistance of Chris Frank. PhastCons scores were computed using the Cistrome conservation plots tool with 1000 bp surrounding the peak center. DHS conservation was further determined by the overlap with mouse CNEEs [120].

The top 50,000 peaks called using F-seq were filtered for intestine-specificity by subtracting the overlapping intervals (10 bp) with the top 50,000 DH sites from liver and kidney. DH sites 10 kb upstream, 10 kb downstream, or within the gene body of genes up or down regulated in the GF ileum [87] were used for motif predictions. *De no*vo motif predictions were performed with MEME-ChIP [309] using default parameters and no background model. Motifs generated by MEME were queried against TRANSFAC and JASPAR databases using TOMTOM [243]. DH and FAIRE sites were linked to nearby genes and analyzed for functional enrichment using GREAT [338].

5.6.6 RNA extraction for RNA-seq

Total RNA was extracted from (i) groups of 6 dpf whole zebrafish larvae from 6 dpf zebrafish (10 larvae per group, 2 biological replicate groups per condition per experiment, 2 experimental replicates total) (ii) IECs from adult zebrafish (iii) IECs from the mouse duodenum, ileum, and colon using TRIzol Reagent (Invitrogen) or the Qiagen RNeasy (Qiagen) kit using manufacturer's protocol. Two μg (in 50 μl RNase-free water) were used for TruSeq library preparation (performed by the UNC High Through-put Sequencing Core) for mRNA Illumina sequencing using 2 x 50bp paired-end reads. Four samples were multiplexed per lane.

5.6 Acknowledgements

I am grateful to Chris Frank for DNase-seq library preparation and computational support; to Lingyun Song for DNase-seq protocol support; to Yoichiro Shibata for help aligning DNase-seq reads and peak calling; to Nick Gomez and Ian Davis for PFG equipment support; to Jeremy Simon for FAIRE-seq protocol advice, aligning FAIRE-seq sequencing reads, and FAIRE-seq peak calling; to Dan McKay and Colin Lickwar for

FAIRE-seq protocol support; to Maureen Bower and Chris Packey for providing GF and CONV-R mice; to Barry Udis and Joan Kalnitsky for flow cytometry; to Jim Notwell and Gill Bejerano for the zCNE dataset; to Adam Gracz and Scott Magness for IEC isolation protocol support; o Agostina Santoro, Sarah Bortvedt, Amanda Mah, and Dallas Donohoe for intestinal tissue processing help; to Praveen Sethupathy for computational advice; and to the labs of Elie Tzima and Jim Faber for equipment support.

CHAPTER 6

Future Prospectus

6.1 Overview

Animals have co-existed with an intestinal microbiota since their inception over half a billion years ago, yet the impact of the intestinal microbiota on vertebrate evolution is not well understood. Much more emphasis has been placed on the microbial communities themselves and on the purported impact of the microbiota on the host as interpreted through the lens of biomedicine. In my thesis work, I have taken a gene-centric and genomic view of transcriptional regulatory landscapes in the vertebrate intestine and made progress in understanding how the microbiota impacts theses landscapes. This prospectus section presents a unique opportunity to propose future trajectories at the interface of vertebrate-microbial co-evolution that can build upon the data presented in this dissertation. I discuss current limitations of the zebrafish model and potential solutions for studying transcriptional regulation, suggest that novel questions can be addressed by extending host-microbe transcriptional genomic research into model vertebrates that have extensive ecological knowledge, and highlight how new waves of vertebrate genome sequences will offer rich resources for those interested in the impact of host-microbe symbiosis on transcriptional regulation.

6.2 Zebrafish and Transcriptional Regulation Analysis

The work throughout my dissertation revolved around using the zebrafish model to study the spatiotemporal regulation of transcription in the context of host-microbiota symbiosis. I elucidated multiple non-genic regulatory regions that appeared to function as discrete modules controlling expression of *angptl4* in distinct tissues including the intestine, liver, and pancreatic islet (Chapter 3). However, we were never able to definitively test this prediction. A clear future direction for our lab, and the field in general, would be to show that deletion of the putative tissue-specific regulatory region at the endogenous locus results in loss or reduction of gene expression in that tissue. It is well known in Drosophila that many genes have redundant (aka "shadow") enhancers that foster robustness to gene regulatory programs [339,340] and it remains unclear if these same concepts will be true for non-developmental gene regulation in the vertebrate intestine. This goal will be best addressed in the short term by BAC recombineering [339,341], but hopefully in the near future efficient methods for homologous recombination of transgenic DNA with genomic DNA or use of TALEN mutagenesis will ready for use the zebrafish.

Another clear direction forward is to transition our understanding of the CRMs controlling transcription of *angptl4* in the zebrafish to mammalian systems. In my work, I showed that the orthologous third intron of mouse *Angptl4* does not drive expression in the liver, islet, or intestine when heterologously assayed in the zebrafish (Figure 3.S8). This is despite numerous reports of a mammalian conserved and functional PPAR response element (PPRE) located in this intron in the mouse that is active in multiple tissues including the liver [177]. Indeed there are PPREs located in the third intron of zebrafish *angptl4* (Figure 3.7). Why then doesn't the mouse region function as a tissue specific enhancer in the zebrafish? I uncovered an intestine-specific DNase hypersensitive site approximately 6.5 kb upstream of the mouse *Angptl4* transcription

start site that is has a similar composition of TFBSs as the zebrafish in3.4 module and is bound by the transcription factor FXR *in vivo* (Figure 5.8). It will be exciting to test this mouse region for heterologous reporter expression in the zebrafish and mammalian intestinal epithelial cells. Futhermore, it would be highly informative to assay the zebrafish in3.3 and in3.4 islet and intestinal regulatory modules in the mouse, as well as other fish models such as medaka or stickleback. Cases in which modules are functional in one organism but not the other organism will yield insight into the logic governing CRM function and co-evolution with corresponding transcription factors. In this light, the mouse and zebrafish DNase-seq and FAIRE-seq datasets will be immense repositories of potential CRMs to study cross-species functional conservation.

I imagine that the utility of the zebrafish for genomics-based assays will increase in the future, especially with the application of genomic methods to uncover non-coding functional DNA. I previously discussed a number of areas that need to be improved in order to maximize the zebrafish as a model, but I will reiterate and compile these points below. Overcoming some of these limitations will further advance our ability use the zebrafish to uncover the transcription factors and cis-regulatory DNA that mediate host-microbe co-evolution.

i. *Limitation:* Most of the conserved non-genic regions between zebrafish and other sequenced vertebrates are involved in development and therefore using conservation as a proxy for function is severely limited to a subset of biological processes [124].

   Potential solution: Sequence the genome of an experimentally tractable fish species that diverged from the zebrafish 10-40 million years ago. Note that there will soon be two more *Danio* species publically available (*Danio nigrofasciatus,*

Danio albolineatus) with genome sequences available (Dave Parichy personal communication), however these species are probably too close to be of much use for discovering non-coding functional regions.

ii. *Limitation:* The Tol2 system offers efficient transgenesis, but does not allow site-specific insertion into the zebrafish genome thus leading to position and copy number effects. Also, the current Tol2 cloning system is hindered by a multi-step cloning procedure.

Potential Solution: Learn from work in Drosophila to develop new strategies such as Cre-recombinase [342], phi31-integrase [343], or Flp-recombinase [344] for site-directed integration of transgenes. For the short-term using the current Tol2 system, my strategy would be to convert the attL gateway site in the pT2cfosGW destination vector into an attP gateway site. This would generate a one-step cloning system for direct cloning of a PCR fragment into a Tol 2 vector upstream of the *cfos* minimal promoter facilitating at least a two-fold increase in throughput.

iii. *Limitation:* The throughput for assaying putative *cis*-regulatory modules for reporter expression *in vivo* is currently only moderate in the zebrafish system.

Potential Solution: There are two strategies that could be employed to increase throughput. The first would be to explore automation of reporter screening [345] because manually aligning, imaging, and scoring injected fish is a major time constraint. Another strategy is to harness high throughput sequencing technologies to massively test putative CRMs in parallel [346,347]. One could

imagine cloning selected FAIRE-seq, DNase-seq, or ChIP-seq sites into one-step cloning vectors upstream of a barcoded reporter, gavaging pooled constructs into the intestine, electroporating vectors into IECs, harvesting RNA after several of days, and sequencing to look for enriched barcodes. Another potentially feasible strategy would be to perform a chromatin assay such as FAIRE- or ChIP-seq, ligate adapters to the DNA pools enriched for regulatory modules, and clone pools into a vector to create a library of CRMs upstream of a barcoded reporter. Sequence the library to identify the CRM and barcoded reporter pair. Then inject libraries into embryos or gavage/electroporate into intestines and sequence reporter RNA.

iv. *Limitation:* A disappointment in my dissertation work was the inability to successfully identify and assay strong candidate transcription factors that regulate the in3.4 CRM using unbiased DNA centered methods. In hindsight, the strength of the zebrafish as a model system is not in biochemistry, but rather in genetic screens. The limitation for my work was a lack of resources to systematically assay transcription factors *in vivo.*

Potential Solution: A large-scale, organized effort aimed at establishing a library of transcription factor resources freely available to the zebrafish community would greatly enhance the discovery power of the zebrafish model. This could begin with a comprehensive effort to assay morpholinos targeting as many TFs as possible and use RNA-seq to identify genes regulated by the TF, generation and validation of ChIP-grade TF-specific antibiodies, vector libraries containing transcription factor TF-specific cDNA for RNA overexpression, Yeast 1-hybrid

TF-specific libraries, or vector libraries of TALEN or Zinc-Finger nucleases targeting TF loci.

v. *Limitation:* I successfully applied FAIRE-seq to uncover the genome-wide regulatory landscape in zebrafish 6 dpf GI tracts. However, this data set includes many cell types and a future directive would be to perform genomic assays on purified cell populations. Current estimate of the number of cells required for assays such as FAIRE-seq and DNase-seq is around 1 million cells per experiment. My attempts to use flow cytometry to sort intestinal epithelial cells and neutrophils recovered approximately 20,000 (IECs) and 3,000 (neutrophils) making the 1 million cell mark out of practical reach.

Potential Solution: It may be possible to optimize these genomic assays on much fewer cells using existing or easily adaptable protocols [348]. Furthermore, microfluidic devices offers incredible opportunities at miniaturization of the many steps in these protocols and are becoming increasingly accessible [349]. Processing tissues and dissociating cells from their *in vivo* context within the living fish may confound results attained from sorted cells. However, for ChIP and FAIRE, it is possible to directly fix the tissue with formaldehyde prior to dissociation and sorting [350]

6.3 Host-Microbe Symbiosis and Adaptive Evolution

I like to imagine the multi-cellular organism as a symbiotic community of unicellular organisms, where each cell has a common yet differentially accessible genome, and a unifying purpose to ensure the propagation of the germ-line. There are

problems with this imagery, but it allows one to invoke principles from ecology and evolution to understand cells as individuals within populations. Recasting the view of the microbiota in ecological terms has already generated significant insight into rules governing microbial community assembly [20] and viewing cells constituting the intestinal epithelium, as well as other cell types within the body, as individual “species” may help us understand the patterns that are emerging at the transcriptional genomics scale. In this light, a major future initiative would be to study how different cell types utilize their accessible genome to maintain homeostasis during changing environmental conditions, such as the presence and absence of a microbiota or in disease states. Are mechanisms distinct for a given cell type or environmental condition? Do mechanisms evolve differentially for different cell types? Intestinal epithelial cells such as enterocytes maintain homeostasis in a dynamic and potentially hostile luminal environment. How do these conditions compare to those of a striated muscle cell or an astrocyte or an olfactory neuron? What impact does cell location and function have on regulatory evolution? Recent evidence suggests that waves of regulatory innovations occurred at specific functional gene categories during evolution [120]. Can this concept be extended to cell type? Cells within a multicellular organism experience very different local environments, so it seems reasonable that homeostatic regulatory networks active in one cell-type may evolve at a different rate than networks in another cell type. A crude analysis in Chapter 5 (Figure 5.6) did not suggest a striking difference in conservation between intestine, lever, and kidney DH sites, however a different pattern may be revealed by binning DH sites into functional categories. Many of these questions are addressable using the immense datasets generated by the ENCODE project as well as recent surveys in mouse [317], however the key insight will come by varying the environmental conditions experienced by the cells and assaying the genome-wide chromatin landscape.

New environments experienced through adaptive radiations would have imposed variable selective pressures on the cell types within an organism, which would have distinct impacts on organism fitness. Organisms such as stickleback [351] or Peromyscus (deer mice) [352] in which there is extensive ecological and evolutionary knowledge, could help us understand how quickly transcriptional regulatory regions can evolve in neutral and adaptive contexts for different cell types. In terms of host-microbe symbiosis, these models represent fertile terrain for exploring how new habitats impact microbial community composition and the resulting host responses.

6.4 Host-Microbe Symbiosis and Genome Sandboxes

In a few years there will be thousands of sequenced genomes and soon thereafter tens of thousands [24]. What can be done with this immense repository of base-pairs? Computation is a tool, much like a microscope, and its creative application to understanding genomes will be a useful experimental skill for anyone interested in host-microbe symbioses. The flood of genomes and the expertise of computational biologists will allow the genetic reconstruction of our evolutionary history and population dynamics. As experimentalists, our role will be to define relevant problems and learn to use computation and genomic resources to tackle them. There are currently 29 mammalian genomes and I can imagine that a high-resolution reconstruction of the genome of the common ancestor of all mammals to be on the horizon [353]. Ancestral genomes will dramatically enrich the ability to understand the role that the intestinal microbiota has played in shaping an animal's physiology. The first step is to catalog the genic and non-genic functional regions of the mouse genome in the intestinal epithelium that are likely to be impacted by the activities of the microbiota (such as in Chapter 5). The next step will be to compare these regions to the ancestral genome. What regions were lost from the ancestral genome or gained, constrained, or accelerated in the

murine lineage that might have impacted gene expression in the mouse intestine? Is there overlap in other extant mammalian lineages such as humans? Evolutionary constraint [121] and lineage-specific losses [354,355] can be calculated with the current set of genomes and we will partially address these questions with our datasets.

However, the most challenging, and naturally interesting, problem will be to determine the causal changes in genomic sequences that underlie phenotypic evolution (or the reciprocal task to understand the conditions that led to changes in functional genomic sequence). For this, a centralized database of well-defined microbiota associated phenotypes (MAPs) including microbial community composition, metagenomics, metabolomics, host gene expression, and physiological impact could be collated for a set of species. There is already a copious amount of phenotypic data in the literature, though my experience has led me to believe that knowledge is fragmented and interpreted with many agendas. Once such a phenotypic database is established, a systematic strategy could be employed to correlate phenotypes with changes in host genotype using statistical analysis of metadata such as maximum information coefficient (MIC) [356] or by mapping MAPs to phylogenies and inferring causal loci ("Forward genomics") [357]. An extended set of gnotobiotic animals would further help define a microbial cause of observed phenotypic traits through cross-species genomic comparisons. It would be great to generate DNase-seq/FAIRE-seq/RNA-seq datasets from one or more species representing both basal and derived branches (with respect to mouse) of the mammalian lineage amenable to gnotobiotic culturing techniques (perhaps rat and guinea pig). Phenotypic and genomic resources for humans are un-paralleled and defining the microbial impact at base-pair resolution on the evolution of human associated traits should be tractable. Defining the causative molecular mechanism underlying a phenotype is one obstacle, proving an adaptive advantage and thus evolutionary significance is a non-trivial next step. With this goal in mind, it would be

fun to extend research of the microbial impact on transcriptional genomics into other mammalian or fish models in which extensive ecological and population genetic knowledge is available, such as those described above.

6.5 Concluding Remarks

Collective human knowledge is a living, breathing organism of gargantuan size with a voracious appetite. I have begun to understand and even appreciate the incremental nature of scientific progress. I realize that the background, data, and interpretations discussed in this dissertation represent little more than a “cell” in the “body” of human knowledge, yet I believe this work will lead to a better understanding of how vertebrate genomes are used to mediate responses to microbes in our intestine. I see potential for this work to influence the future directions of research into host-microbe symbiosis, and I sincerely hope that insights from this data can be further developed to positively impact society either through biomedicine or scientific education. At the very least, this process has satisfied some of my curiosity. These are exciting times in science and new methods are rapidly extending the realm of tractable questions that can be addressed in model genetic organisms as well as non-model systems. The future promises space and time for these ideas to grow, new cells to be discovered, and projected utility to become realized.

REFERENCES

1. Bonner JT (1997) The Origins of Multicellularity. Int Biol 1: 27-36.

2. King N, Westbrook MJ, Young SL, Kuo A, Abedin M, et al. (2008) The genome of the choanoflagellate Monosiga brevicollis and the origin of metazoans. Nature 451: 783-788.

3. Hooper LV, Midtvedt T, Gordon JI (2002) How host-microbial interactions shape the nutrient environment of the mammalian intestine. Annu Rev Nutr 22: 283-307.

4. Clayton TA, Baker D, Lindon JC, Everett JR, Nicholson JK (2009) Pharmacometabonomic identification of a significant host-microbiome metabolic interaction affecting human drug metabolism. Proc Natl Acad Sci U S A 106: 14728-14733.

5. Wei X, Yang Z, Rey FE, Ridaura VK, Davidson NO, et al. (2012) Fatty acid synthase modulates intestinal barrier function through palmitoylation of mucin 2. Cell Host Microbe 11: 140-152.

6. Lee YK, Mazmanian SK (2010) Has the microbiota played a critical role in the evolution of the adaptive immune system? Science 330: 1768-1773.

7. Hooper LV, Littman DR, Macpherson AJ (2012) Interactions between the microbiota and the immune system. Science 336: 1268-1273.

8. Backhed F, Ding H, Wang T, Hooper LV, Koh GY, et al. (2004) The gut microbiota as an environmental factor that regulates fat storage. Proc Natl Acad Sci U S A 101: 15718-15723.

9. Ley RE, Turnbaugh PJ, Klein S, Gordon JI (2006) Microbial ecology: human gut microbes associated with obesity. Nature 444: 1022-1023.

10. Turnbaugh PJ, Ley RE, Mahowald MA, Magrini V, Mardis ER, et al. (2006) An obesity-associated gut microbiome with increased capacity for energy harvest. Nature 444: 1027-1031.

11. Backhed F, Manchester JK, Semenkovich CF, Gordon JI (2007) Mechanisms underlying the resistance to diet-induced obesity in germ-free mice. Proc Natl Acad Sci U S A 104: 979-984.

12. Vijay-Kumar M, Aitken JD, Carvalho FA, Cullender TC, Mwangi S, et al. (2010) Metabolic syndrome and altered gut microbiota in mice lacking Toll-like receptor 5. Science 328: 228-231.

13. Wen L, Ley RE, Volchkov PY, Stranges PB, Avanesyan L, et al. (2008) Innate immunity and intestinal microbiota in the development of Type 1 diabetes. Nature 455: 1109-1113.

14. Mazmanian SK, Round JL, Kasper DL (2008) A microbial symbiosis factor prevents intestinal inflammatory disease. Nature 453: 620-625.

15. Sartor RB (2008) Microbial influences in inflammatory bowel diseases. Gastroenterology 134: 577-594.

16. Wang Z, Klipfell E, Bennett BJ, Koeth R, Levison BS, et al. (2011) Gut flora metabolism of phosphatidylcholine promotes cardiovascular disease. Nature 472: 57-63.

17. Blumberg R, Powrie F (2012) Microbiota, disease, and back to health: a metastable journey. Sci Transl Med 4: 137rv137.

18. Dethlefsen L, McFall-Ngai M, Relman DA (2007) An ecological and evolutionary perspective on human-microbe mutualism and disease. Nature 449: 811-818.

19. Little AE, Robinson CJ, Peterson SB, Raffa KF, Handelsman J (2008) Rules of engagement: interspecies interactions that regulate microbial communities. Annu Rev Microbiol 62: 375-401.

20. Camp JG, Kanther M, Semova I, Rawls JF (2009) Patterns and scales in gastrointestinal microbial ecology. Gastroenterology 136: 1989-2002.

21. Costello EK, Stagaman K, Dethlefsen L, Bohannan BJ, Relman DA (2012) The application of ecological theory toward an understanding of the human microbiome. Science 336: 1255-1262.

22. Shendure J, Ji H (2008) Next-generation DNA sequencing. Nat Biotechnol 26: 1135-1145.

23. Mardis ER (2008) Next-generation DNA sequencing methods. Annu Rev Genomics Hum Genet 9: 387-402.

24. Zerbino DR, Paten B, Haussler D (2012) Integrating genomes. Science 336: 179-182.

25. Jimenez E, Delgado S, Fernandez L, Garcia N, Albujar M, et al. (2008) Assessment of the bacterial diversity of human colostrum and screening of staphylococcal and enterococcal populations for potential virulence factors. Res Microbiol 159: 595-601.

26. Benton JM (1998) The quality of the fossil record of vertebrates; Donovan SK, Paul, C. R. C., editor. New York: Wiley.

27. Woese CR, Fox GE (1977) Phylogenetic structure of the prokaryotic domain: the primary kingdoms. Proc Natl Acad Sci U S A 74: 5088-5090.

28. Lane DJ, Pace B, Olsen GJ, Stahl DA, Sogin ML, et al. (1985) Rapid determination of 16S ribosomal RNA sequences for phylogenetic analyses. Proc Natl Acad Sci U S A 82: 6955-6959.

29. Eckburg PB, Bik EM, Bernstein CN, Purdom E, Dethlefsen L, et al. (2005) Diversity of the human intestinal microbial flora. Science 308: 1635-1638.

30. Turnbaugh PJ, Hamady M, Yatsunenko T, Cantarel BL, Duncan A, et al. (2009) A core gut microbiome in obese and lean twins. Nature 457: 480-484.

31. Ley RE, Hamady M, Lozupone C, Turnbaugh PJ, Ramey RR, et al. (2008) Evolution of mammals and their gut microbes. Science 320: 1647-1651.

32. Roeselers G, Mittge EK, Stephens WZ, Parichy DM, Cavanaugh CM, et al. (2011) Evidence for a core gut microbiota in the zebrafish. ISME J 5: 1595-1608.

33. Turnbaugh PJ, Ley RE, Hamady M, Fraser-Liggett CM, Knight R, et al. (2007) The human microbiome project. Nature 449: 804-810.

34. HMPC (2012) Structure, function and diversity of the healthy human microbiome. Nature 486: 207-214.

35. Fukami T, Beaumont HJ, Zhang XX, Rainey PB (2007) Immigration history controls diversification in experimental adaptive radiation. Nature 446: 436-439.

36. Spor A, Koren O, Ley R (2011) Unravelling the effects of the environment and host genotype on the gut microbiome. Nat Rev Microbiol 9: 279-290.

37. Backhed F, Ley RE, Sonnenburg JL, Peterson DA, Gordon JI (2005) Host-bacterial mutualism in the human intestine. Science 307: 1915-1920.

38. Handelsman J, Rondon MR, Brady SF, Clardy J, Goodman RM (1998) Molecular biological access to the chemistry of unknown soil microbes: a new frontier for natural products. Chem Biol 5: R245-249.

39. Yatsunenko T, Rey FE, Manary MJ, Trehan I, Dominguez-Bello MG, et al. (2012) Human gut microbiome viewed across age and geography. Nature 486: 222-227.

40. Ley RE, Lozupone CA, Hamady M, Knight R, Gordon JI (2008) Worlds within worlds: evolution of the vertebrate gut microbiota. Nat Rev Microbiol 6: 776-788.

41. Turnbaugh PJ, Backhed F, Fulton L, Gordon JI (2008) Diet-induced obesity is linked to marked but reversible alterations in the mouse distal gut microbiome. Cell Host Microbe 3: 213-223.

42. Nutrition SoLA (1995) Nutrient Requirements of Laboratory Animals. Washington: National Academy Press.

43. Lawrence C (2007) The husbandry of zebrafish (Danio rerio): A review. Aquaculture 269: 1-20.

44. Wilson RP (2002) Amino Acids and Proteins; Halver JEH, R. W., editor. San Diego: Academic Press.

45. Stevens CEaH, I. D. (1995) Comparative Physiology of the Vertebrate Digestive System. Cambridge: Cambridge University Press.

46. Boron WFaB, E. L. (2009) Medical Physiology. Philadelphia: Saunders Elsevier.

47. Wallace KN, Akhter S, Smith EM, Lorent K, Pack M (2005) Intestinal growth and differentiation in zebrafish. Mech Dev 122: 157-173.

48. van der Flier LG, Clevers H (2009) Stem cells, self-renewal, and differentiation in the intestinal epithelium. Annu Rev Physiol 71: 241-260.

49. Oehlers SH, Flores MV, Chen T, Hall CJ, Crosier KE, et al. (2011) Topographical distribution of antimicrobial genes in the zebrafish intestine. Dev Comp Immunol 35: 385-391.

50. Schonhoff SE, Giel-Moloney M, Leiter AB (2004) Minireview: Development and differentiation of gut endocrine cells. Endocrinology 145: 2639-2644.

51. Gill SR, Pop M, Deboy RT, Eckburg PB, Turnbaugh PJ, et al. (2006) Metagenomic analysis of the human distal gut microbiome. Science 312: 1355-1359.

52. Qin J, Li R, Raes J, Arumugam M, Burgdorf KS, et al. (2010) A human gut microbial gene catalogue established by metagenomic sequencing. Nature 464: 59-65.

53. Nicholson JK, Holmes E, Kinross J, Burcelin R, Gibson G, et al. (2012) Host-gut microbiota metabolic interactions. Science 336: 1262-1267.

54. Nuttal GHFaT, H. (1895) Thierisches Leben ohne Bacterien im Verdauungskanal. Z Physiol Chem 21: 109-121.

55. Reyniers JA, Trexler, P. C., Ervin, R. F. (1946) Rearing germfree albino rats; Reyniers JA, editor. Notre Dame, Ind: Univ. Notre Dame.

56. Coates MEaGBE (1984) The Germ-free Animal in Biomedical Research. London: Royal Society of Medicine Press.

57. Pham LN, Kanther M, Semova I, Rawls JF (2008) Methods for generating and colonizing gnotobiotic zebrafish. Nat Protoc 3: 1862-1875.

58. Wostmann BS (1996) Germfree and gnotobiotic models: background and applications. Boca Raton, FL: CRC Press.

59. Midtvedt T (1974) Microbial bile acid transformation. Am J Clin Nutr 27: 1341-1347.

60. Saxerholt H, Midtvedt T (1986) Intestinal deconjugation of bilirubin in germfree and conventional rats. Scand J Clin Lab Invest 46: 341-344.

61. Salyers AA, West SE, Vercellotti JR, Wilkins TD (1977) Fermentation of mucins and plant polysaccharides by anaerobic bacteria from the human colon. Appl Environ Microbiol 34: 529-533.

62. Savage DC, Siegel JE, Snellen JE, Whitt DD (1981) Transit-Time of Epithelial-Cells in the Small-Intestines of Germ-Free Mice and Ex-Germfree Mice Associated with Indigenous Microorganisms. Applied and Environmental Microbiology 42: 996-1001.

63. Turnbaugh PJ, Gordon JI (2008) An invitation to the marriage of metagenomics and metabolomics. Cell 134: 708-713.

64. Semova I, Carten, J.D., Stombaugh, J., Mackey, L.C., Knight, R., Farber, S.A., Rawls, J.F. (2012) Microbiota and diet regulate fatty acid absorption in the zebrafish intestine. Cell Host Microbe.

65. (2004) The ENCODE (ENCyclopedia Of DNA Elements) Project. Science 306: 636-640.

66. Celniker SE, Dillon LA, Gerstein MB, Gunsalus KC, Henikoff S, et al. (2009) Unlocking the secrets of the genome. Nature 459: 927-930.

67. Birney E, Stamatoyannopoulos JA, Dutta A, Guigo R, Gingeras TR, et al. (2007) Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. Nature 447: 799-816.

68. Visel A, Rubin EM, Pennacchio LA (2009) Genomic views of distant-acting enhancers. Nature 461: 199-205.

69. Roy S, Ernst J, Kharchenko PV, Kheradpour P, Negre N, et al. (2010) Identification of functional elements and regulatory circuits by Drosophila modENCODE. Science 330: 1787-1797.

70. Gerstein MB, Lu ZJ, Van Nostrand EL, Cheng C, Arshinoff BI, et al. (2010) Integrative analysis of the Caenorhabditis elegans genome by the modENCODE project. Science 330: 1775-1787.

71. Hurwitz J (2005) The discovery of RNA polymerase. J Biol Chem 280: 42477-42485.

72. Jacob F, Monod J (1961) Genetic Regulatory Mechanisms in Synthesis of Proteins. Journal of Molecular Biology 3: 318-&.

73. Crick F (1970) Central Dogma of Molecular Biology. Nature 227: 561-&.

74. Taatjes DJ (2010) The human Mediator complex: a versatile, genome-wide regulator of transcription. Trends Biochem Sci 35: 315-322.

75. Buratowski S (2009) Progression through the RNA polymerase II CTD cycle. Mol Cell 36: 541-546.

76. Margaritis T, Holstege FC (2008) Poised RNA polymerase II gives pause for thought. Cell 133: 581-584.

77. Li B, Carey M, Workman JL (2007) The role of chromatin during transcription. Cell 128: 707-719.

78. Clapier CR, Cairns BR (2009) The biology of chromatin remodeling complexes. Annu Rev Biochem 78: 273-304.

79. Jacquier A (2009) The complex eukaryotic transcriptome: unexpected pervasive transcription and novel small RNAs. Nat Rev Genet 10: 833-844.

80. Mercer TR, Dinger ME, Mattick JS (2009) Long non-coding RNAs: insights into functions. Nat Rev Genet 10: 155-159.

81. Bustamante C, Cheng W, Mejia YX (2011) Revisiting the central dogma one molecule at a time. Cell 144: 480-497.

82. Raj A, van Oudenaarden A (2008) Nature, nurture, or chance: stochastic gene expression and its consequences. Cell 135: 216-226.

83. Maston GA, Evans SK, Green MR (2006) Transcriptional regulatory elements in the human genome. Annual Review of Genomics and Human Genetics 7: 29-59.

84. Panne D, Maniatis T, Harrison SC (2007) An atomic model of the interferon-beta enhanceosome. Cell 129: 1111-1123.

85. Bulger M, Groudine M (2011) Functional and mechanistic diversity of distal transcription enhancers. Cell 144: 327-339.

86. Kowalczyk MS, Hughes JR, Garrick D, Lynch MD, Sharpe JA, et al. (2012) Intragenic enhancers act as alternative promoters. Mol Cell 45: 447-458.

87. Larsson E, Tremaroli V, Lee YS, Koren O, Nookaew I, et al. (2012) Analysis of gut microbial regulation of host gene expression along the length of the gut and regulation of gut microbial ecology through MyD88. Gut 61: 1124-1131.

88. El Aidy S, Merrifield CA, Derrien M, van Baarlen P, Hooiveld G, et al. (2012) The gut microbiota elicits a profound metabolic reorientation in the mouse jejunal mucosa during conventionalisation. Gut.

89. El Aidy S, van Baarlen P, Derrien M, Lindenbergh-Kortleve DJ, Hooiveld G, et al. (2012) Temporal and spatial interplay of microbiota and intestinal mucosa drive establishment of immune homeostasis in conventionalized mice. Mucosal Immunol.

90. Rawls JF, Samuel BS, Gordon JI (2004) Gnotobiotic zebrafish reveal evolutionarily conserved responses to the gut microbiota. Proc Natl Acad Sci U S A 101: 4596-4601.

91. Chung H, Pamp SJ, Hill JA, Surana NK, Edelman SM, et al. (2012) Gut immune maturation depends on colonization with a host-specific microbiota. Cell 149: 1578-1593.

92. Maloy KJ, Powrie F (2011) Intestinal homeostasis and its breakdown in inflammatory bowel disease. Nature 474: 298-306.

93. Duerkop BA, Vaishnava S, Hooper LV (2009) Immune responses to the microbiota at the intestinal mucosal surface. Immunity 31: 368-376.

94. Kanther M, Sun X, Muhlbauer M, Mackey LC, Flynn EJ, 3rd, et al. (2011) Microbial Colonization Induces Dynamic Temporal and Spatial Patterns of NF-kappaB Activation in the Zebrafish Digestive Tract. Gastroenterology 141: 197-207.

95. Hayden MS, Ghosh S (2012) NF-kappaB, the first quarter-century: remarkable progress and outstanding questions. Genes Dev 26: 203-234.

96. Battle MA, Bondow BJ, Iverson MA, Adams SJ, Jandacek RJ, et al. (2008) GATA4 is essential for jejunal function in mice. Gastroenterology 135: 1676-1686 e1671.

97. Dusing MR, Wiginton DA (2005) Epithelial lineages of the small intestine have unique patterns of GATA expression. J Mol Histol 36: 15-24.

98. Beuling E, Baffour-Awuah NY, Stapleton KA, Aronson BE, Noah TK, et al. (2011) GATA factors regulate proliferation, differentiation, and gene expression in small intestine of mature mice. Gastroenterology 140: 1219-1229 e1211-1212.

99. Thisse C, Thisse B (2008) High-resolution in situ hybridization to whole-mount zebrafish embryos. Nat Protoc 3: 59-69.

100. Zeng L, Carter AD, Childs SJ (2009) miR-145 directs intestinal maturation in zebrafish. Proc Natl Acad Sci U S A 106: 17793-17798.

101. Shulzhenko N, Morgun A, Hsiao W, Battle M, Yao M, et al. (2011) Crosstalk between B lymphocytes, microbiota and the intestinal epithelium governs immunity versus metabolism in the gut. Nat Med 17: 1585-1593.

102. Shapira M, Hamlin BJ, Rong J, Chen K, Ronen M, et al. (2006) A conserved role for a GATA transcription factor in regulating epithelial innate immune responses. Proc Natl Acad Sci U S A 103: 14086-14091.

103. Chinetti G, Fruchart JC, Staels B (2000) Peroxisome proliferator-activated receptors (PPARs): nuclear receptors at the crossroads between lipid metabolism and inflammation. Inflamm Res 49: 497-505.

104. Shih DQ, Bussen M, Sehayek E, Ananthanarayanan M, Shneider BL, et al. (2001) Hepatocyte nuclear factor-1alpha is an essential regulator of bile acid and plasma cholesterol metabolism. Nat Genet 27: 375-382.

105. Mitro N, Mak PA, Vargas L, Godio C, Hampton E, et al. (2007) The nuclear receptor LXR is a glucose sensor. Nature 445: 219-223.

106. Peet DJ, Turley SD, Ma W, Janowski BA, Lobaccaro JM, et al. (1998) Cholesterol and bile acid metabolism are impaired in mice lacking the nuclear oxysterol receptor LXR alpha. Cell 93: 693-704.

107. Makishima M, Lu TT, Xie W, Whitfield GK, Domoto H, et al. (2002) Vitamin D receptor as an intestinal bile acid sensor. Science 296: 1313-1316.

108. Sinal CJ, Tohkin M, Miyata M, Ward JM, Lambert G, et al. (2000) Targeted disruption of the nuclear receptor FXR/BAR impairs bile acid and lipid homeostasis. Cell 102: 731-744.

109. Van der Flier LG, Sabates-Bellver J, Oving I, Haegebarth A, De Palo M, et al. (2007) The Intestinal Wnt/TCF Signature. Gastroenterology 132: 628-632.

110. Silberg DG, Swain GP, Suh ER, Traber PG (2000) Cdx1 and cdx2 expression during intestinal development. Gastroenterology 119: 961-971.

111. Blache P, van de Wetering M, Duluc I, Domon C, Berta P, et al. (2004) SOX9 is an intestine crypt transcription factor, is regulated by the Wnt pathway, and represses the CDX2 and MUC2 genes. J Cell Biol 166: 37-47.

112. Crosnier C, Stamataki D, Lewis J (2006) Organizing cell renewal in the intestine: stem cells, signals and combinatorial control. Nat Rev Genet 7: 349-359.

113. Ghaleb AM, Aggarwal G, Bialkowska AB, Nandan MO, Yang VW (2008) Notch inhibits expression of the Kruppel-like factor 4 tumor suppressor in the intestinal epithelium. Mol Cancer Res 6: 1920-1927.

114. Sancho E, Batlle E, Clevers H (2004) Signaling pathways in intestinal development and cancer. Annu Rev Cell Dev Biol 20: 695-723.

115. Theiss AL, Fruchtman S, Lund PK (2004) Growth factors in inflammatory bowel disease: the actions and interactions of growth hormone and insulin-like growth factor-I. Inflamm Bowel Dis 10: 871-880.

116. Hardison RC, Taylor J (2012) Genomic approaches towards finding cis-regulatory modules in animals. Nat Rev Genet 13: 469-483.

117. (2009) Genome 10K: a proposal to obtain whole-genome sequence for 10,000 vertebrate species. J Hered 100: 659-674.

118. (2010) A map of human genome variation from population-scale sequencing. Nature 467: 1061-1073.

119. Bejerano G, Pheasant M, Makunin I, Stephen S, Kent WJ, et al. (2004) Ultraconserved elements in the human genome. Science 304: 1321-1325.

120. Lowe CB, Kellis M, Siepel A, Raney BJ, Clamp M, et al. (2011) Three periods of regulatory innovation during vertebrate evolution. Science 333: 1019-1024.

121. Lindblad-Toh K, Garber M, Zuk O, Lin MF, Parker BJ, et al. (2011) A high-resolution map of human evolutionary constraint using 29 mammals. Nature 478: 476-482.

122. Katzman S, Kern AD, Bejerano G, Fewell G, Fulton L, et al. (2007) Human genome ultraconserved elements are ultraselected. Science 317: 915.

123. Pennacchio LA, Ahituv N, Moses AM, Prabhakar S, Nobrega MA, et al. (2006) In vivo enhancer analysis of human conserved non-coding sequences. Nature 444: 499-502.

124. Woolfe A, Goodson M, Goode DK, Snell P, McEwen GK, et al. (2005) Highly conserved non-coding sequences are associated with vertebrate development. PLoS Biol 3: e7.

125. Luger K, Mader AW, Richmond RK, Sargent DF, Richmond TJ (1997) Crystal structure of the nucleosome core particle at 2.8 A resolution. Nature 389: 251-260.

126. Jiang C, Pugh BF (2009) Nucleosome positioning and gene regulation: advances through genomics. Nat Rev Genet 10: 161-172.

127. Schones DE, Zhao K (2008) Genome-wide approaches to studying chromatin modifications. Nat Rev Genet 9: 179-191.

128. Tan M, Luo H, Lee S, Jin F, Yang JS, et al. (2011) Identification of 67 histone marks and histone lysine crotonylation as a new type of histone modification. Cell 146: 1016-1028.

129. Park PJ (2009) ChIP-seq: advantages and challenges of a maturing technology. Nat Rev Genet 10: 669-680.

130. Heintzman ND, Hon GC, Hawkins RD, Kheradpour P, Stark A, et al. (2009) Histone modifications at human enhancers reflect global cell-type-specific gene expression. Nature 459: 108-112.

131. Creyghton MP, Cheng AW, Welstead GG, Kooistra T, Carey BW, et al. (2010) Histone H3K27ac separates active from poised enhancers and predicts developmental state. Proc Natl Acad Sci U S A 107: 21931-21936.

132. Aday AW, Zhu LJ, Lakshmanan A, Wang J, Lawson ND (2011) Identification of cis regulatory features in the embryonic zebrafish genome through large-scale profiling of H3K4me1 and H3K4me3 binding sites. Dev Biol 357: 450-462.

133. Vastenhouw NL, Zhang Y, Woods IG, Imam F, Regev A, et al. (2010) Chromatin signature of embryonic pluripotency is established during genome activation. Nature 464: 922-926.

134. Lickwar CR, Mueller F, Hanlon SE, McNally JG, Lieb JD (2012) Genome-wide protein-DNA binding dynamics suggest a molecular clutch for transcription factor function. Nature 484: 251-255.

135. Rhee HS, Pugh BF (2011) Comprehensive genome-wide protein-DNA interactions detected at single-nucleotide resolution. Cell 147: 1408-1419.

136. Boyle AP, Davis S, Shulha HP, Meltzer P, Margulies EH, et al. (2008) High-resolution mapping and characterization of open chromatin across the genome. Cell 132: 311-322.

137. Song L, Crawford GE (2010) DNase-seq: a high-resolution technique for mapping active gene regulatory elements across the genome from mammalian cells. Cold Spring Harb Protoc 2010: pdb prot5384.

138. Giresi PG, Kim J, McDaniell RM, Iyer VR, Lieb JD (2007) FAIRE (Formaldehyde-Assisted Isolation of Regulatory Elements) isolates active regulatory elements from human chromatin. Genome Res 17: 877-885.

139. Simon JM, Giresi PG, Davis IJ, Lieb JD (2012) Using formaldehyde-assisted isolation of regulatory elements (FAIRE) to isolate active regulatory DNA. Nat Protoc 7: 256-267.

140. Galas DJ, Schmitz A (1978) DNAse footprinting: a simple method for the detection of protein-DNA binding specificity. Nucleic Acids Res 5: 3157-3170.

141. Song L, Zhang Z, Grasfeder LL, Boyle AP, Giresi PG, et al. (2011) Open chromatin defined by DNaseI and FAIRE identifies regulatory elements that shape cell-type identity. Genome Res 21: 1757-1767.

142. Boyle AP, Song L, Lee BK, London D, Keefe D, et al. (2011) High-resolution genome-wide in vivo footprinting of diverse transcription factors in human cells. Genome Res 21: 456-464.

143. Verzi MP, Shin H, He HH, Sulahian R, Meyer CA, et al. (2010) Differentiation-specific histone modifications reveal dynamic chromatin interactions and partners for the intestinal transcription factor CDX2. Dev Cell 19: 713-726.

144. Verzi MP, Shin H, Ho LL, Liu XS, Shivdasani RA (2011) Essential and redundant functions of caudal family proteins in activating adult intestinal genes. Mol Cell Biol 31: 2026-2039.

145. Thomas AM, Hart SN, Kong B, Fang J, Zhong XB, et al. (2010) Genome-wide tissue-specific farnesoid X receptor binding in mouse liver and intestine. Hepatology 51: 1410-1419.

146. Akhtar-Zaidi B, Cowper-Sal-lari R, Corradin O, Saiakhova A, Bartels CF, et al. (2012) Epigenomic enhancer profiling defines a signature of colon cancer. Science 336: 736-739.

147. May D, Blow MJ, Kaplan T, McCulley DJ, Jensen BC, et al. (2012) Large-scale discovery of enhancers from human heart tissue. Nat Genet 44: 89-93.

148. Liberati NT, Urbach JM, Miyata S, Lee DG, Drenkard E, et al. (2006) An ordered, nonredundant library of Pseudomonas aeruginosa strain PA14 transposon insertion mutants. Proc Natl Acad Sci U S A 103: 2833-2838.

149. Peterson RT, Fishman MC (2011) Designing zebrafish chemical screens. Methods Cell Biol 105: 525-541.

150. Laggner C, Kokel D, Setola V, Tolia A, Lin H, et al. (2012) Chemical informatics and target identification in a zebrafish phenotypic screen. Nat Chem Biol 8: 144-146.

151. Pedron T, Sansonetti P (2008) Commensals, bacterial pathogens and intestinal inflammation: an intriguing menage a trois. Cell Host Microbe 3: 344-347.

152. Kleinjan DA, Lettice LA (2008) Long-range gene control and genetic disease. Adv Genet 61: 339-388.

153. Lee JC, Parkes M (2011) Genome-wide association studies and Crohn's disease. Brief Funct Genomics 10: 71-76.

154. Gallo RL, Hooper LV (2012) Epithelial antimicrobial defence of the skin and intestine. Nat Rev Immunol 12: 503-516.

155. Wray GA (2007) The evolutionary significance of cis-regulatory mutations. Nat Rev Genet 8: 206-216.

156. Carroll SB (2008) Evo-devo and an expanding evolutionary synthesis: a genetic theory of morphological evolution. Cell 134: 25-36.

157. Chan YF, Marks ME, Jones FC, Villarreal G, Jr., Shapiro MD, et al. (2010) Adaptive evolution of pelvic reduction in sticklebacks by recurrent deletion of a Pitx1 enhancer. Science 327: 302-305.

158. Noonan JP (2009) Regulatory DNAs and the evolution of human development. Curr Opin Genet Dev 19: 557-564.

159. Flint HJ, Bayer EA, Rincon MT, Lamed R, White BA (2008) Polysaccharide utilization by gut bacteria: potential for new insights from genomic analysis. Nat Rev Microbiol 6: 121-131.

160. Wang H, Eckel RH (2009) Lipoprotein lipase: from gene to obesity. Am J Physiol Endocrinol Metab 297: E271-288.

161. Fleissner CK, Huebel N, Abd El-Bary MM, Loh G, Klaus S, et al. (2010) Absence of intestinal microbiota does not protect mice from diet-induced obesity. Br J Nutr 104: 919-929.

162. Koster A, Chao YB, Mosior M, Ford A, Gonzalez-DeWhitt PA, et al. (2005) Transgenic angiopoietin-like (angptl)4 overexpression and targeted disruption of angptl4 and angptl3: regulation of triglyceride metabolism. Endocrinology 146: 4943-4950.

163. Desai U, Lee EC, Chung K, Gao C, Gay J, et al. (2007) Lipid-lowering effects of anti-angiopoietin-like 4 antibody recapitulate the lipid phenotype found in angiopoietin-like 4 knockout mice. Proc Natl Acad Sci U S A 104: 11766-11771.

164. Lee EC, Desai U, Gololobov G, Hong S, Feng X, et al. (2009) Identification of a new functional domain in angiopoietin-like 3 (ANGPTL3) and angiopoietin-like 4 (ANGPTL4) involved in binding and inhibition of lipoprotein lipase (LPL). J Biol Chem 284: 13735-13745.

165. Yau MH, Wang Y, Lam KS, Zhang J, Wu D, et al. (2009) A highly conserved motif within the NH2-terminal coiled-coil domain of angiopoietin-like protein 4 confers its inhibitory effects on lipoprotein lipase by disrupting the enzyme dimerization. J Biol Chem 284: 11942-11952.

166. Romeo S, Pennacchio LA, Fu Y, Boerwinkle E, Tybjaerg-Hansen A, et al. (2007) Population-based resequencing of ANGPTL4 uncovers variations that reduce triglycerides and increase HDL. Nat Genet 39: 513-516.

167. Folsom AR, Peacock JM, Demerath E, Boerwinkle E (2008) Variation in ANGPTL4 and risk of coronary heart disease: the Atherosclerosis Risk in Communities Study. Metabolism 57: 1591-1596.

168. Cazes A, Galaup A, Chomel C, Bignon M, Brechot N, et al. (2006) Extracellular matrix-bound angiopoietin-like 4 inhibits endothelial cell adhesion, migration, and sprouting and alters actin cytoskeleton. Circ Res 99: 1207-1215.

169. Zhu P, Tan MJ, Huang RL, Tan CK, Chong HC, et al. (2011) Angiopoietin-like 4 protein elevates the prosurvival intracellular O2(-):H2O2 ratio and confers anoikis resistance to tumors. Cancer Cell 19: 401-415.

170. Padua D, Zhang XH, Wang Q, Nadal C, Gerald WL, et al. (2008) TGFbeta primes breast tumors for lung metastasis seeding through angiopoietin-like 4. Cell 133: 66-77.

171. Galaup A, Cazes A, Le Jan S, Philippe J, Connault E, et al. (2006) Angiopoietin-like 4 prevents metastasis through inhibition of vascular permeability and tumor cell motility and invasiveness. Proc Natl Acad Sci U S A 103: 18721-18726.

172. Goh YY, Pal M, Chong HC, Zhu P, Tan MJ, et al. (2010) Angiopoietin-like 4 interacts with matrix proteins to modulate wound healing. J Biol Chem 285: 32999-33009.

173. Kersten S, Mandard S, Tan NS, Escher P, Metzger D, et al. (2000) Characterization of the fasting-induced adipose factor FIAF, a novel peroxisome proliferator-activated receptor target gene. J Biol Chem 275: 28488-28493.

174. Yoon JC, Chickering TW, Rosen ED, Dussault B, Qin Y, et al. (2000) Peroxisome proliferator-activated receptor gamma target gene encoding a novel angiopoietin-related protein associated with adipose differentiation. Mol Cell Biol 20: 5343-5349.

175. Kutlu B, Burdick D, Baxter D, Rasschaert J, Flamez D, et al. (2009) Detailed transcriptome atlas of the pancreatic beta cell. BMC Med Genomics 2: 3.

176. Bikopoulos G, da Silva Pimenta A, Lee SC, Lakey JR, Der SD, et al. (2008) Ex vivo transcriptional profiling of human pancreatic islets following chronic exposure to monounsaturated fatty acids. J Endocrinol 196: 455-464.

177. Mandard S, Zandbergen F, Tan NS, Escher P, Patsouris D, et al. (2004) The direct peroxisome proliferator-activated receptor target fasting-induced adipose factor (FIAF/PGAR/ANGPTL4) is present in blood plasma as a truncated protein that is increased by fenofibrate treatment. J Biol Chem 279: 34411-34420.

178. Staiger H, Haas C, Machann J, Werner R, Weisser M, et al. (2009) Muscle-derived angiopoietin-like protein 4 is induced by fatty acids via peroxisome proliferator-activated receptor (PPAR)-delta and is of metabolic relevance in humans. Diabetes 58: 579-589.

179. Georgiadi A, Lichtenstein L, Degenhardt T, Boekschoten MV, van Bilsen M, et al. (2010) Induction of cardiac Angptl4 by dietary fatty acids is mediated by peroxisome proliferator-activated receptor beta/delta and protects against fatty acid-induced oxidative stress. Circ Res 106: 1712-1721.

180. Kaddatz K, Adhikary T, Finkernagel F, Meissner W, Muller-Brusselbach S, et al. (2010) Transcriptional profiling identifies functional interactions of TGF beta and PPAR beta/delta signaling: synergistic induction of ANGPTL4 transcription. J Biol Chem 285: 29469-29479.

181. Aronsson L, Huang Y, Parini P, Korach-Andre M, Hakansson J, et al. (2010) Decreased fat storage by Lactobacillus paracasei is associated with increased levels of angiopoietin-like 4 protein (ANGPTL4). PLoS One 5.

182. Koliwad SK, Kuo T, Shipp LE, Gray NE, Backhed F, et al. (2009) Angiopoietin-like 4 (ANGPTL4, fasting-induced adipose factor) is a direct glucocorticoid receptor target and participates in glucocorticoid-regulated triglyceride metabolism. J Biol Chem 284: 25593-25601.

183. Belanger AJ, Lu H, Date T, Liu LX, Vincent KA, et al. (2002) Hypoxia up-regulates expression of peroxisome proliferator-activated receptor gamma angiopoietin-related gene (PGAR) in cardiomyocytes: role of hypoxia inducible factor 1alpha. J Mol Cell Cardiol 34: 765-774.

184. Wang B, Wood IS, Trayhurn P (2007) Dysregulation of the expression and secretion of inflammation-related adipokines by hypoxia in human adipocytes. Pflugers Arch 455: 479-492.

185. Kawakami K (2004) Transgenesis and gene trap methods in zebrafish by using the Tol2 transposable element. Methods Cell Biol 77: 201-222.

186. Fisher S, Grice EA, Vinton RM, Bessling SL, Urasaki A, et al. (2006) Evaluating the biological relevance of putative enhancers using Tol2 transposon-mediated transgenesis in zebrafish. Nat Protoc 1: 1297-1305.

187. Pack M, Solnica-Krezel L, Malicki J, Neuhauss SC, Schier AF, et al. (1996) Mutations affecting development of zebrafish digestive organs. Development 123: 321-328.

188. Field HA, Dong PD, Beis D, Stainier DY (2003) Formation of the digestive system in zebrafish. II. Pancreas morphogenesis. Dev Biol 261: 197-208.

189. Chu J, Sadler KC (2009) New school in liver development: lessons from zebrafish. Hepatology 50: 1656-1663.

190. Ng AN, de Jong-Curtain TA, Mawdsley DJ, White SJ, Shin J, et al. (2005) Formation of the digestive system in zebrafish: III. Intestinal epithelium morphogenesis. Dev Biol 286: 114-135.

191. Bates JM, Mittge E, Kuhlman J, Baden KN, Cheesman SE, et al. (2006) Distinct signals from the microbiota promote different aspects of zebrafish gut differentiation. Dev Biol 297: 374-386.

192. Rawls JF, Mahowald MA, Goodman AL, Trent CM, Gordon JI (2007) In vivo imaging and genetic analysis link bacterial motility and symbiosis in the zebrafish gut. Proc Natl Acad Sci U S A 104: 7622-7627.

193. Hama K, Provost E, Baranowski TC, Rubinstein AL, Anderson JL, et al. (2009) In vivo imaging of zebrafish digestive organ function using multiple quenched fluorescent reporters. Am J Physiol Gastrointest Liver Physiol 296: G445-453.

194. Milligan-Myhre K, Charette JR, Phennicie RT, Stephens WZ, Rawls JF, et al. (2011) Study of host-microbe interactions in zebrafish. Methods Cell Biol 105: 87-116.

195. Rawls JF, Mahowald MA, Ley RE, Gordon JI (2006) Reciprocal gut microbiota transplants from zebrafish and mice to germ-free recipients reveal host habitat selection. Cell 127: 423-433.

196. Fisher S, Grice EA, Vinton RM, Bessling SL, McCallion AS (2006) Conservation of RET regulatory function from human to zebrafish without sequence similarity. Science 312: 276-279.

197. Navratilova P, Fredman D, Hawkins TA, Turner K, Lenhard B, et al. (2009) Systematic human/zebrafish comparative identification of cis-regulatory activity around vertebrate developmental transcription factor genes. Dev Biol 327: 526-540.

198. Tsujimura T, Hosoya T, Kawamura S (2010) A single enhancer regulating the differential expression of duplicated red-sensitive opsin genes in zebrafish. PLoS Genet 6: e1001245.

199. Chao CH, Wang HD, Yuh CH (2010) Complexity of cis-regulatory organization of six3a during forebrain and eye development in zebrafish. BMC Dev Biol 10: 35.

200. Ng CE, Yokomizo T, Yamashita N, Cirovic B, Jin H, et al. (2010) A Runx1 intronic enhancer marks hemogenic endothelial cells and hematopoietic stem cells. Stem Cells 28: 1869-1881.

201. Borok MJ, Tran DA, Ho MC, Drewell RA (2010) Dissecting the regulatory switches of development: lessons from enhancer evolution in Drosophila. Development 137: 5-13.

202. Wasserman WW, Sandelin A (2004) Applied bioinformatics for the identification of regulatory elements. Nat Rev Genet 5: 276-287.

203. Haeussler M, Joly JS (2011) When needles look like hay: how to find tissue-specific enhancers in model organism genomes. Dev Biol 350: 239-254.

204. Hedges S (2009) Vertebrates; Hedges SB KS, editor: Oxford University Press.

205. Peng ZD, R. He, S. (2009) Teleost fishes.; Hedges SB KS, editor: Oxford University Press.

206. Curado S, Anderson RM, Jungblut B, Mumm J, Schroeter E, et al. (2007) Conditional targeted cell ablation in zebrafish: a new tool for regeneration studies. Dev Dyn 236: 1025-1035.

207. Hare EE, Peterson BK, Iyer VN, Meier R, Eisen MB (2008) Sepsid even-skipped enhancers are functionally conserved in Drosophila despite lack of sequence conservation. PLoS Genet 4: e1000106.

208. Tang KL, Agnew MK, Hirt MV, Sado T, Schneider LM, et al. (2010) Systematics of the subfamily Danioninae (Teleostei: Cypriniformes: Cyprinidae). Mol Phylogenet Evol 57: 189-214.

209. Quigley IK, Manuel JL, Roberts RA, Nuckels RJ, Herrington ER, et al. (2005) Evolutionary diversification of pigment pattern in Danio fishes: differential fms dependence and stripe loss in D. albolineatus. Development 132: 89-104.

210. Jonas JC, Laybutt DR, Steil GM, Trivedi N, Pertusa JG, et al. (2001) High glucose stimulates early response gene c-Myc expression in rat pancreatic beta cells. J Biol Chem 276: 35375-35381.

211. Pascal SM, Guiot Y, Pelengaris S, Khan M, Jonas JC (2008) Effects of c-MYC activation on glucose stimulus-secretion coupling events in mouse pancreatic islets. Am J Physiol Endocrinol Metab 295: E92-102.

212. Gunton JE, Kulkarni RN, Yim S, Okada T, Hawthorne WJ, et al. (2005) Loss of ARNT/HIF1beta mediates altered gene expression and pancreatic-islet dysfunction in human type 2 diabetes. Cell 122: 337-349.

213. Pillai R, Huypens P, Huang M, Schaefer S, Sheinin T, et al. (2011) Aryl hydrocarbon receptor nuclear translocator/hypoxia-inducible factor-1{beta} plays a critical role in maintaining glucose-stimulated anaplerosis and insulin release from pancreatic {beta}-cells. J Biol Chem 286: 1014-1024.

214. Martin CC, Svitek CA, Oeser JK, Henderson E, Stein R, et al. (2003) Upstream stimulatory factor (USF) and neurogenic differentiation/beta-cell E box transactivator 2 (NeuroD/BETA2) contribute to islet-specific glucose-6-phosphatase catalytic-subunit-related protein (IGRP) gene expression. Biochem J 371: 675-686.

215. Han SI, Yasuda K, Kataoka K (2011) ATF2 interacts with beta-cell-enriched transcription factors, MafA, Pdx1, and beta2, and activates insulin gene transcription. J Biol Chem 286: 10449-10456.

216. Semple CA, Gautier P, Taylor K, Dorin JR (2006) The changing of the guard: Molecular diversity and rapid evolution of beta-defensins. Mol Divers 10: 575-584.

217. Levine M (2010) Transcriptional enhancers in animal development and evolution. Curr Biol 20: R754-763.

218. Meireles-Filho AC, Stark A (2009) Comparative genomics of gene regulation-conservation and divergence of cis-regulatory information. Curr Opin Genet Dev 19: 565-570.

219. Stark A, Lin MF, Kheradpour P, Pedersen JS, Parts L, et al. (2007) Discovery of functional elements in 12 Drosophila genomes using evolutionary signatures. Nature 450: 219-232.

220. Hooper LV, Wong MH, Thelin A, Hansson L, Falk PG, et al. (2001) Molecular analysis of commensal host-microbial relationships in the intestine. Science 291: 881-884.

221. Grootaert C, Van de Wiele T, Van Roosbroeck I, Possemiers S, Vercoutter-Edouart AS, et al. (2011) Bacterial monocultures, propionate, butyrate and H(2) O(2) modulate the expression, secretion and structure of the fasting-induced adipose factor in gut epithelial cell lines. Environ Microbiol 13: 1778-1789.

222. Davidson E (2006) The Regulatory Genome: Gene Regulatory Networks In Development And Evolution: Academic Press. 304 p.

223. Palanker L, Tennessen JM, Lam G, Thummel CS (2009) Drosophila HNF4 regulates lipid mobilization and beta-oxidation. Cell Metab 9: 228-239.

224. Hayhurst GP, Lee YH, Lambert G, Ward JM, Gonzalez FJ (2001) Hepatocyte nuclear factor 4alpha (nuclear receptor 2A1) is essential for maintenance of hepatic gene expression and lipid homeostasis. Mol Cell Biol 21: 1393-1403.

225. Thisse B, Pfumio, S., Fürthauer, M., Loppin B., Heyer, V., Degrave, A., Woehl, R., Lux, A., Steffan, T., Charbonnier, X.Q. and Thisse, C (2001) Expression of the zebrafish genome during embryogenesis. ZFIN on-line publication.

226. Flynn EJ, 3rd, Trent CM, Rawls JF (2009) Ontogeny and nutritional control of adipogenesis in zebrafish (Danio rerio). J Lipid Res 50: 1641-1652.

227. Bertrand S, Thisse B, Tavares R, Sachs L, Chaumot A, et al. (2007) Unexpected novel relational links uncovered by extensive developmental profiling of nuclear receptor expression. PLoS Genet 3: e188.

228. Heinaniemi M, Uski JO, Degenhardt T, Carlberg C (2007) Meta-analysis of primary target genes of peroxisome proliferator-activated receptors. Genome Biol 8: R147.

229. Edgar RC (2004) MUSCLE: a multiple sequence alignment method with reduced time and space complexity. BMC Bioinformatics 5: 113.

230. Lei X, Shi F, Basu D, Huq A, Routhier S, et al. (2011) Proteolytic processing of angiopoietin-like protein 4 by proprotein convertases modulates its inhibitory effects on lipoprotein lipase activity. J Biol Chem 286: 15747-15756.

231. Letunic I, Doerks T, Bork P (2009) SMART 6: recent updates and new developments. Nucleic Acids Res 37: D229-232.

232. Finn RD, Mistry J, Tate J, Coggill P, Heger A, et al. (2010) The Pfam protein families database. Nucleic Acids Res 38: D211-222.

233. Gasteiger E, Gattiker A, Hoogland C, Ivanyi I, Appel RD, et al. (2003) ExPASy: The proteomics server for in-depth protein knowledge and analysis. Nucleic Acids Res 31: 3784-3788.

234. Felsenstein J (2005) Using the quantitative genetic threshold model for inferences between and within species. Philos Trans R Soc Lond B Biol Sci 360: 1427-1434.

235. Brudno M, Do CB, Cooper GM, Kim MF, Davydov E, et al. (2003) LAGAN and Multi-LAGAN: efficient tools for large-scale multiple alignment of genomic DNA. Genome Res 13: 721-731.

236. Frazer KA, Pachter L, Poliakov A, Rubin EM, Dubchak I (2004) VISTA: computational tools for comparative genomics. Nucleic Acids Res 32: W273-279.

237. Matys V, Kel-Margoulis OV, Fricke E, Liebich I, Land S, et al. (2006) TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes. Nucleic Acids Res 34: D108-110.

238. Bryne JC, Valen E, Tang MH, Marstrand T, Winther O, et al. (2008) JASPAR, the open access database of transcription factor-binding profiles: new content and tools in the 2008 update. Nucleic Acids Res 36: D102-106.

239. Chekmenev DS, Haid C, Kel AE (2005) P-Match: transcription factor binding site search by combining patterns and weight matrices. Nucleic Acids Res 33: W432-437.

240. Schug J, Overton, GC (1998) TESS: Transcription Element Search Software on the WWW. http://wwwcbilupennedu/cgi-bin/tess/tess?RQ=WELCOME.

241. Bailey TL, Elkan C (1994) Fitting a mixture model by expectation maximization to discover motifs in biopolymers. Proc Int Conf Intell Syst Mol Biol 2: 28-36.

242. Bailey TL, Gribskov M (1998) Combining evidence using p-values: application to sequence homology searches. Bioinformatics 14: 48-54.

243. Gupta S, Stamatoyannopoulos JA, Bailey TL, Noble WS (2007) Quantifying similarity between motifs. Genome Biol 8: R24.

244. Abramoff MD, Magalhaes, P.J., Ram, S.J (2004) Image Processing with ImageJ. Biophotonics International 11: pp. 36-42.

245. Gilbert W, Muller-Hill B (1966) Isolation of the lac repressor. Proc Natl Acad Sci U S A 56: 1891-1898.

246. Ptashne M (1967) ISOLATION OF THE lambda PHAGE REPRESSOR. Proc Natl Acad Sci U S A 57: 306-313.

247. Ptashne M (1988) How eukaryotic transcriptional activators work. Nature 335: 683-689.

248. Englesberg E, Irr J, Power J, Lee N (1965) Positive control of enzyme synthesis by gene C in the L-arabinose system. J Bacteriol 90: 946-957.

249. Emmer M, deCrombrugghe B, Pastan I, Perlman R (1970) Cyclic AMP receptor protein of E. coli: its role in the synthesis of inducible enzymes. Proc Natl Acad Sci U S A 66: 480-487.

250. Tjian R (1978) The binding site on SV40 DNA for a T antigen-related protein. Cell 13: 165-179.

251. Engelke DR, Ng SY, Shastry BS, Roeder RG (1980) Specific interaction of a purified transcription factor with an internal control region of 5S RNA genes. Cell 19: 717-728.

252. Payvar F, Wrange O, Carlstedt-Duke J, Okret S, Gustafsson JA, et al. (1981) Purified glucocorticoid receptors bind selectively in vitro to a cloned DNA fragment whose transcription is regulated by glucocorticoids in vivo. Proc Natl Acad Sci U S A 78: 6628-6632.

253. McKnight SL, Kingsbury R (1982) Transcriptional control signals of a eukaryotic protein-coding gene. Science 217: 316-324.

254. Dynan WS, Tjian R (1983) The promoter-specific transcription factor Sp1 binds to upstream sequences in the SV40 early promoter. Cell 35: 79-87.

255. Garvie CW, Wolberger C (2001) Recognition of specific DNA sequences. Mol Cell 8: 937-946.

256. Luscombe NM, Austin SE, Berman HM, Thornton JM (2000) An overview of the structures of protein-DNA complexes. Genome Biol 1: REVIEWS001.

257. Brivanlou AH, Darnell JE, Jr. (2002) Signal transduction and the control of gene expression. Science 295: 813-818.

258. Brent R, Ptashne M (1985) A eukaryotic transcriptional activator bearing the DNA specificity of a prokaryotic repressor. Cell 43: 729-736.

259. Seeman NC, Rosenberg JM, Rich A (1976) Sequence-specific recognition of double helical nucleic acids by proteins. Proc Natl Acad Sci U S A 73: 804-808.

260. Carey MF, Smale, S. T., Peterson C. L. (2008) Transcriptional Regulation in Eukaryotes: Concepts, Strategies, and Techniques. Cold Spring Harbor: Cold Spring Harbor Library Press.

261. Joshi R, Passner JM, Rohs R, Jain R, Sosinsky A, et al. (2007) Functional specificity of a Hox protein mediated by the recognition of minor groove structure. Cell 131: 530-543.

262. Parker SC, Hansen L, Abaan HO, Tullius TD, Margulies EH (2009) Local DNA topography correlates with functional noncoding regions of the human genome. Science 324: 389-392.

263. Vaquerizas JM, Kummerfeld SK, Teichmann SA, Luscombe NM (2009) A census of human transcription factors: function, expression and evolution. Nat Rev Genet 10: 252-263.

264. Segal E, Raveh-Sadka T, Schroeder M, Unnerstall U, Gaul U (2008) Predicting expression patterns from regulatory sequence in Drosophila segmentation. Nature 451: 535-540.

265. De Val S, Chi NC, Meadows SM, Minovitsky S, Anderson JP, et al. (2008) Combinatorial regulation of endothelial gene expression by ets and forkhead transcription factors. Cell 135: 1053-1064.

266. Zinzen RP, Girardot C, Gagneur J, Braun M, Furlong EE (2009) Combinatorial binding predicts spatio-temporal cis-regulatory activity. Nature 462: 65-70.

267. von Hippel PH (2007) From "simple" DNA-protein interactions to the macromolecular machines of gene expression. Annu Rev Biophys Biomol Struct 36: 79-105.

268. Simicevic J, Deplancke B (2010) DNA-centered approaches to characterize regulatory protein-DNA interaction complexes. Mol Biosyst 6: 462-468.

269. Hannenhalli S (2008) Eukaryotic transcription factor binding sites--modeling and integrative search methods. Bioinformatics 24: 1325-1331.

270. Dynan WS, Tjian R (1983) Isolation of transcription factors that discriminate between different promoters recognized by RNA polymerase II. Cell 32: 669-680.

271. Briggs MR, Kadonaga JT, Bell SP, Tjian R (1986) Purification and biochemical characterization of the promoter-specific transcription factor, Sp1. Science 234: 47-52.

272. Kadonaga JT (2004) Regulation of RNA polymerase II transcription by sequence-specific DNA binding factors. Cell 116: 247-257.

273. Wolters DA, Washburn MP, Yates JR, 3rd (2001) An automated multidimensional protein identification technology for shotgun proteomics. Anal Chem 73: 5683-5690.

274. Fournier ML, Gilmore JM, Martin-Brown SA, Washburn MP (2007) Multidimensional separations-based shotgun proteomics. Chem Rev 107: 3654-3686.

275. Jeong Y, Leskow FC, El-Jaick K, Roessler E, Muenke M, et al. (2008) Regulation of a remote Shh forebrain enhancer by the Six3 homeoprotein. Nat Genet 40: 1348-1353.

276. Reed DE, Huang XM, Wohlschlegel JA, Levine MS, Senger K (2008) DEAF-1 regulates immunity gene expression in Drosophila. Proc Natl Acad Sci U S A 105: 8351-8356.

277. Gunter TE, Gunter KK, Sheu SS, Gavin CE (1994) Mitochondrial calcium transport: physiological and pathological relevance. Am J Physiol 267: C313-339.

278. Schreiber E, Matthias P, Muller MM, Schaffner W (1989) Rapid detection of octamer binding proteins with 'mini-extracts', prepared from a small number of cells. Nucleic Acids Res 17: 6419.

279. Carey MF, Peterson CL, Smale ST (2009) Dignam and Roeder nuclear extract preparation. Cold Spring Harb Protoc 2009: pdb prot5330.

280. Lacks SA (1981) Deoxyribonuclease I in mammalian tissues. Specificity of inhibition by actin. J Biol Chem 256: 2644-2648.

281. Reece-Hoyes JS, Barutcu AR, McCord RP, Jeong JS, Jiang L, et al. (2011) Yeast one-hybrid assays for gene-centered human gene regulatory network mapping. Nat Methods 8: 1050-1052.

282. Hens K, Feuz JD, Isakova A, Iagovitina A, Massouras A, et al. (2011) Automated protein-DNA interaction screening of Drosophila regulatory elements. Nat Methods 8: 1065-1070.

283. Reece-Hoyes JS, Diallo A, Lajoie B, Kent A, Shrestha S, et al. (2011) Enhanced yeast one-hybrid assays for high-throughput gene-centered regulatory network mapping. Nat Methods 8: 1059-1064.

284. Badis G, Berger MF, Philippakis AA, Talukder S, Gehrke AR, et al. (2009) Diversity and complexity in DNA recognition by transcription factors. Science 324: 1720-1723.

285. Morcos PA, Li Y, Jiang S (2008) Vivo-Morpholinos: a non-peptide transporter delivers Morpholinos into a wide array of mouse tissues. Biotechniques 45: 613-614, 616, 618 passim.

286. Meng X, Noyes MB, Zhu LJ, Lawson ND, Wolfe SA (2008) Targeted gene inactivation in zebrafish using engineered zinc-finger nucleases. Nat Biotechnol 26: 695-701.

287. McCammon JM, Doyon Y, Amacher SL (2011) Inducing high rates of targeted mutagenesis in zebrafish using zinc finger nucleases (ZFNs). Methods Mol Biol 770: 505-527.

288. Huang P, Xiao A, Zhou M, Zhu Z, Lin S, et al. (2011) Heritable gene targeting in zebrafish using customized TALENs. Nat Biotechnol 29: 699-700.

289. Sander JD, Cade L, Khayter C, Reyon D, Peterson RT, et al. (2011) Targeted gene disruption in somatic zebrafish cells using engineered TALENs. Nat Biotechnol 29: 697-698.

290. Camp JG, Jazwa AL, Trent CM, Rawls JF (2012) Intronic cis-regulatory modules mediate tissue-specific and microbial control of angptl4/fiaf transcription. PLoS Genet 8: e1002585.

291. Gracz AD, Puthoff BJ, Magness ST (2012) Identification, isolation, and culture of intestinal epithelial stem cells from murine intestine. Methods Mol Biol 879: 89-107.

292. Wang Z, Du J, Lam SH, Mathavan S, Matsudaira P, et al. (2010) Morphological and molecular evidence for functional organization along the rostrocaudal axis of the adult zebrafish intestine. BMC Genomics 11: 392.

293. Bradford MM (1976) A rapid and sensitive method for the quantitation of microgram quantities of protein utilizing the principle of protein-dye binding. Anal Biochem 72: 248-254.

294. Li M, Wang B, Zhang M, Rantalainen M, Wang S, et al. (2008) Symbiotic gut microbes modulate human metabolic phenotypes. Proc Natl Acad Sci U S A 105: 2117-2122.

295. Elson CO, Cong Y, McCracken VJ, Dimmitt RA, Lorenz RG, et al. (2005) Experimental models of inflammatory bowel disease reveal innate, adaptive, and regulatory mechanisms of host dialogue with the microbiota. Immunol Rev 206: 260-276.

296. Round JL, Mazmanian SK (2009) The gut microbiota shapes intestinal immune responses during health and disease. Nat Rev Immunol 9: 313-323.

297. Britten RJ, Davidson EH (1969) Gene regulation for higher cells: a theory. Science 165: 349-357.

298. Madison BB, Dunbar L, Qiao XT, Braunstein K, Braunstein E, et al. (2002) Cis elements of the villin gene control expression in restricted domains of the vertical (crypt) and horizontal (duodenum, cecum) axes of the intestine. J Biol Chem 277: 33275-33283.

299. Cheesman SE, Neal JT, Mittge E, Seredick BM, Guillemin K (2011) Epithelial cell proliferation in the developing zebrafish intestine is regulated by the Wnt pathway and microbial signaling via Myd88. Proc Natl Acad Sci U S A 108 Suppl 1: 4570-4577.

300. Parzen E (1962) Estimation of a Probability Density-Function and Mode. Annals of Mathematical Statistics 33: 1065-&.

301. Boyle AP, Guinney J, Crawford GE, Furey TS (2008) F-Seq: a feature density estimator for high-throughput sequence tags. Bioinformatics 24: 2537-2538.

302. Hesselberth JR, Chen X, Zhang Z, Sabo PJ, Sandstrom R, et al. (2009) Global mapping of protein-DNA interactions in vivo by digital genomic footprinting. Nat Methods 6: 283-289.

303. Siepel A, Bejerano G, Pedersen JS, Hinrichs AS, Hou M, et al. (2005) Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. Genome Res 15: 1034-1050.

304. Hinnebusch BF, Siddique A, Henderson JW, Malo MS, Zhang W, et al. (2004) Enterocyte differentiation marker intestinal alkaline phosphatase is a target gene of the gut-enriched Kruppel-like factor. Am J Physiol Gastrointest Liver Physiol 286: G23-30.

305. Sweetser DA, Birkenmeier EH, Klisak IJ, Zollman S, Sparkes RS, et al. (1987) The human and rodent intestinal fatty acid binding protein genes. A comparative analysis of their structure, expression, and linkage relationships. J Biol Chem 262: 16060-16071.

306. Lay JM, Bane G, Brunkan CS, Davis J, Lopez-Diaz L, et al. (2005) Enteroendocrine cell expression of a cholecystokinin gene construct in transgenic mice and cultured cells. Am J Physiol Gastrointest Liver Physiol 288: G354-361.

307. Kirkland SC, Henderson K (2001) Collagen IV synthesis is restricted to the enteroendocrine pathway during multilineage differentiation of human colorectal epithelial stem cells. J Cell Sci 114: 2055-2064.

308. Gum JR, Jr., Hicks JW, Gillespie AM, Carlson EJ, Komuves L, et al. (1999) Goblet cell-specific expression mediated by the MUC2 mucin gene promoter in the intestine of transgenic mice. Am J Physiol 276: G666-676.

309. Machanick P, Bailey TL (2011) MEME-ChIP: motif analysis of large DNA datasets. Bioinformatics 27: 1696-1697.

310. Donohoe DR, Garge N, Zhang X, Sun W, O'Connell TM, et al. (2011) The microbiome and butyrate regulate energy metabolism and autophagy in the mammalian colon. Cell Metab 13: 517-526.

311. Hiller MA, S. Notwell, J. H. Wenger, A. M. Bejerano, G. (in review) Genomic resources annotating the zebrafish cis-regulatory landscape.

312. Matsuzaka T, Shimano H, Yahagi N, Kato T, Atsumi A, et al. (2007) Crucial role of a long-chain fatty acid elongase, Elovl6, in obesity-induced insulin resistance. Nat Med 13: 1193-1202.

313. Mariadason JM, Nicholas C, L'Italien KE, Zhuang M, Smartt HJ, et al. (2005) Gene expression profiling of intestinal epithelial cell maturation along the crypt-villus axis. Gastroenterology 128: 1081-1088.

314. Sato T, van Es JH, Snippert HJ, Stange DE, Vries RG, et al. (2011) Paneth cells constitute the niche for Lgr5 stem cells in intestinal crypts. Nature 469: 415-418.

315. Zhang Y, Liu T, Meyer CA, Eeckhoute J, Johnson DS, et al. (2008) Model-based analysis of ChIP-Seq (MACS). Genome Biol 9: R137.

316. Li G, Ruan X, Auerbach RK, Sandhu KS, Zheng M, et al. (2012) Extensive promoter-centered chromatin interactions provide a topological basis for transcription regulation. Cell 148: 84-98.

317. Shen Y, Yue F, McCleary DF, Ye Z, Edsall L, et al. (2012) A map of the cis-regulatory sequences in the mouse genome. Nature.

318. Heinz S, Benner C, Spann N, Bertolino E, Lin YC, et al. (2010) Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. Mol Cell 38: 576-589.

319. Sharov AA, Ko MS (2009) Exhaustive search for over-represented DNA sequence motifs with CisFinder. DNA Res 16: 261-273.

320. Zhang Z, Chang CW, Goh WL, Sung WK, Cheung E (2011) CENTDIST: discovery of co-associated factors by motif distribution. Nucleic Acids Res 39: W391-399.

321. Tompa M, Li N, Bailey TL, Church GM, De Moor B, et al. (2005) Assessing computational tools for the discovery of transcription factor binding sites. Nat Biotechnol 23: 137-144.

322. Chan ET, Quon GT, Chua G, Babak T, Trochesset M, et al. (2009) Conservation of core gene expression in vertebrate tissues. J Biol 8: 33.

323. Weirauch MT, Hughes TR (2010) Conserved expression without conserved regulatory sequence: the more things change, the more they stay the same. Trends Genet 26: 66-74.

324. Schmidt D, Wilson MD, Ballester B, Schwalie PC, Brown GD, et al. (2010) Five-vertebrate ChIP-seq reveals the evolutionary dynamics of transcription factor binding. Science 328: 1036-1040.

325. Fu C, Wehr DR, Edwards J, Hauge B (2008) Rapid one-step recombinational cloning. Nucleic Acids Res 36: e54.

326. Pashos EE, Kague E, Fisher S (2008) Evaluation of cis-regulatory function in zebrafish. Brief Funct Genomic Proteomic 7: 465-473.

327. Abud HE, Lock P, Heath JK (2004) Efficient gene transfer into the epithelial cell layer of embryonic mouse intestine using low-voltage electroporation. Gastroenterology 126: 1779-1787.

328. Koo BK, Stange DE, Sato T, Karthaus W, Farin HF, et al. (2012) Controlled gene expression in primary Lgr5 organoid cultures. Nat Methods 9: 81-83.

329. Shibata Y, Sheffield NC, Fedrigo O, Babbitt CC, Wortham M, et al. (2012) Extensive Evolutionary Changes in Regulatory Element Activity during Human Origins Are Associated with Altered Gene Expression and Positive Selection. PLoS Genet 8: e1002789.

330. Blanchette M, Bataille AR, Chen X, Poitras C, Laganiere J, et al. (2006) Genome-wide computational prediction of transcriptional regulatory modules reveals new insights into human gene expression. Genome Res 16: 656-668.

331. Degner JF, Pai AA, Pique-Regi R, Veyrieras JB, Gaffney DJ, et al. (2012) DNase I sensitivity QTLs are a major determinant of human expression variation. Nature 482: 390-394.

332. Wu W, Cheng Y, Keller CA, Ernst J, Kumar SA, et al. (2011) Dynamics of the epigenetic landscape during erythroid differentiation after GATA1 restoration. Genome Res 21: 1659-1671.

333. Li H, Durbin R (2010) Fast and accurate long-read alignment with Burrows-Wheeler transform. Bioinformatics 26: 589-595.

334. Rashid NU, Giresi PG, Ibrahim JG, Sun W, Lieb JD (2011) ZINBA integrates local covariates with DNA-seq data to identify broad and narrow regions of enrichment, even within amplified genomic regions. Genome Biol 12: R67.

335. Goecks J, Nekrutenko A, Taylor J (2010) Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. Genome Biol 11: R86.

336. Giardine B, Riemer C, Hardison RC, Burhans R, Elnitski L, et al. (2005) Galaxy: a platform for interactive large-scale genome analysis. Genome Res 15: 1451-1455.

337. Liu T, Ortiz JA, Taing L, Meyer CA, Lee B, et al. (2011) Cistrome: an integrative platform for transcriptional regulation studies. Genome Biol 12: R83.

338. McLean CY, Bristor D, Hiller M, Clarke SL, Schaar BT, et al. (2010) GREAT improves functional interpretation of cis-regulatory regions. Nat Biotechnol 28: 495-501.

339. Perry MW, Boettiger AN, Bothma JP, Levine M (2010) Shadow enhancers foster robustness of Drosophila gastrulation. Curr Biol 20: 1562-1567.

340. Hong JW, Hendrix DA, Levine MS (2008) Shadow enhancers as a source of evolutionary novelty. Science 321: 1314.

341. Bussmann J, Schulte-Merker S (2011) Rapid BAC selection for tol2-mediated transgenesis in zebrafish. Development 138: 4327-4332.

342. Oberstein A, Pare A, Kaplan L, Small S (2005) Site-specific transgenesis by Cre-mediated recombination in Drosophila. Nat Methods 2: 583-585.

343. Fish MP, Groth AC, Calos MP, Nusse R (2007) Creating transgenic Drosophila by microinjecting the site-specific phiC31 integrase mRNA and a transgene-containing donor plasmid. Nat Protoc 2: 2325-2331.

344. Venken KJ, Bellen HJ (2012) Genome-wide manipulations of Drosophila melanogaster with transposons, Flp recombinase, and PhiC31 integrase. Methods Mol Biol 859: 203-228.

345. Peravali R, Gehrig J, Giselbrecht S, Lutjohann DS, Hadzhiev Y, et al. (2011) Automated feature detection and imaging for high-resolution screening of zebrafish embryos. Biotechniques 50: 319-324.

346. Melnikov A, Murugan A, Zhang X, Tesileanu T, Wang L, et al. (2012) Systematic dissection and optimization of inducible enhancers in human cells using a massively parallel reporter assay. Nat Biotechnol 30: 271-277.

347. Patwardhan RP, Hiatt JB, Witten DM, Kim MJ, Smith RP, et al. (2012) Massively parallel functional dissection of mammalian enhancers in vivo. Nat Biotechnol 30: 265-270.

348. Xu CR, Cole PA, Meyers DJ, Kormish J, Dent S, et al. (2011) Chromatin "prepattern" and histone modifiers in a fate choice for liver and pancreas. Science 332: 963-966.

349. Thorsen T, Maerkl SJ, Quake SR (2002) Microfluidic large-scale integration. Science 298: 580-584.

350. Bonn S, Zinzen RP, Perez-Gonzalez A, Riddell A, Gavin AC, et al. (2012) Cell type-specific chromatin immunoprecipitation from multicellular complex samples using BiTS-ChIP. Nat Protoc 7: 978-994.

351. Jones FC, Grabherr MG, Chan YF, Russell P, Mauceli E, et al. (2012) The genomic basis of adaptive evolution in threespine sticklebacks. Nature 484: 55-61.

352. Manceau M, Domingues VS, Mallarino R, Hoekstra HE (2011) The developmental role of Agouti in color pattern evolution. Science 331: 1062-1065.

353. Blanchette M, Green ED, Miller W, Haussler D (2004) Reconstructing large regions of an ancestral mammalian genome in silico. Genome Res 14: 2412-2423.

354. Hiller M, Schaar, B.T., Bejerano, G. (2012) Hundreds of conserved non-coding genomic regions are independently lost in mammals. submitted.

355. McLean CY, Reno PL, Pollen AA, Bassan AI, Capellini TD, et al. (2011) Human-specific loss of regulatory DNA and the evolution of human-specific traits. Nature 471: 216-219.

356. Reshef DN, Reshef YA, Finucane HK, Grossman SR, McVean G, et al. (2011) Detecting novel associations in large data sets. Science 334: 1518-1524.

357. Hiller M, Schaar, B.T., Indjeian, V.B., Kingsley, D.T., Hagey, L.R., Bejerano, G. (2012) Forward genomics links genotype to phenotype using independent phenotypic losses among related species. submitted.