From Genome to Morphology: The Dissection of a Developmental Gene Regulatory Network

by

Tanvi Shashikant

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy

Department of Biological Sciences
Carnegie Mellon University

April 21st 2017

Thesis Advisor: Dr. Charles A Ettensohn

To my husband Abhay Prasanna,
My parents Dr. Uma Shashikant and H.R. Shashikant,
My brother Prithvi Shashikant,
You are my world.

Acknowledgements

First, I want to thank my advisor Dr. Charles Ettensohn for his support, encouragement, guidance and mentorship. Chuck provided an intellectually stimulating lab environment for me to discover my research interests, and the support and confidence to enable me to pursue them with abandon. I learned not only how to do science, but also how to think like a scientist. Chuck cared about my overall development as a person, and pushed me to pursue additional learning opportunities as well as a fulfilling personal life. He saw me for who I was, and honed me to become who I am today: for that, I will always be grateful.

I want to thank my thesis committee: Dr. Javier Lopez, Dr. Joel C McManus and Dr. Deborah Chapman for pushing me to be more ambitious, for being invested in my goals and progress, and for providing a space where all manner of scientific ideas and discussions can emerge, only to be transformed into tangible and achievable research plans. Thank you for always shining light on the ultimate goal, the bigger picture, to keep me focussed and motivated.

Ettensohn Lab members, past and present, have been instrumental to my happiness and well-being in lab. They have gone above and beyond to help me learn and troubleshoot experiments, dissect and understand critical developmental biology concepts, and navigate the challenges of graduate school in a country far from home. Dr. Ashrifia Adomako-Ankomah was my first mentor who incited in me a passion and excitement for all things sea urchin. Dr. Kiran Rafiq inspired me to think genome-wide, and to go boldly where no Ettensohnian has gone before. Dr. Zhongling Sun motivated me to work harder. Jian Ming Khor uplifts me with his passion for science and his enthusiasm for my work. Dr. Debleena Dey and Nathalie Chen provided enthusiastic Molecular Biology support.

I am thankful for the guidance and support I received from the Hinman Lab (especially Dr. Gregory Cary), McManus Lab and Mitchell Lab.

I am very grateful for funding received from the National Science Foundation and the Department of Biological Sciences that made this work possible.

The Department of Biological Sciences has provided me immense support and a sense of community. I want to especially thank Emily Stark, Ena Miceli, Nathan Urban, Shoba Subramanian and the Mellon Storeroom.

My friends and batchmates who celebrated and commiserated with me: thank you for being there.

Contents

List of Figures v
List of Tables vi
List of Supplementary Figures vi
Abstract 1

1 Introduction **3**
1.1 An Introduction to Gene Regulatory Networks (GRNs) 3
1.2 Using the Sea Urchin Embryo to Study GRNs 5
1.3 Skeletogenesis in the Sea Urchin Embryo 6
1.4 The Primary Mesenchyme Cell GRN 7
1.4.1 Initial Activation of the Network in the Skeletogenic Lineage 7
1.4.2 Activation of Early and Late Regulatory Genes 9
1.4.3 Activation of Skeletal Effector Genes 11
1.5 Using the PMC GRN to Study the Evolution of Developmental Programs . 13
1.5.1 Skeletogensis in Cidaroids 15
1.5.2 Skeletogenesis in Holothuroids 16
1.5.3 The Ophiuroid Skeletogenic Program 17
1.5.4 Mesoderm Specification in Asteroids 17
1.5.5 The Evolution of the Larval Endoskeleton in Echinoderms: Insights from Comparative GRN Analysis 18
1.6 Conclusions 19
Bibliography 21

2 Genome-wide Identification of Skeletal Morphogenesis Genes **28**
2.1 Abstract 28
2.2 Introduction 29
2.3 Results 30
2.3.1 RNA-seq analysis of mRNAs differentially expressed by PMCs at the onset of gastrulation 30
2.3.2 Characterization of DE genes 33
2.3.3 Transcriptional inputs into DE genes 38
2.3.4 Non-DE genes 40
2.4 Discussion 42

2.4.1 The identification of morphogenetic effector genes 42
2.4.2 Regulatory inputs into effector genes 43
2.4.3 MAPK Signaling and the PMC GRN 44
2.4.4 The evolution of biomineralization 45
2.5 Materials and Methods 46
2.6 Supplementary Figures 49
Bibliography 52

3 Chromatin Accessibility Profiling Identifies *Cis*-regulatory Modules in an Early Embryonic Cell Lineage 59
3.1 Abstract 59
3.2 Background 60
3.3 Results 62
3.3.1 Analysis of U0126-dependent, hypersensitive sites identified by DNase-seq 62
3.3.2 Analysis of differentially hypersensitive sites in PMCs identified by ATAC-seq 66
3.3.3 Correspondence between DNase-seq and ATAC-seq datasets 69
3.3.4 Differential chromatin accessibility mapping identifies known PMC CRMs 70
3.3.5 Validation of newly discovered PMC CRMs using GFP reporter gene assays 71
3.3.6 Computational prediction of transcription factor binding sites in PMC CRMs 74
3.4 Discussion 76
3.5 Conclusions 80
3.6 Methods 81
3.7 Supplementary Figures 86
Bibliography 90

4 Conclusions and Future Directions 97

List of Figures

1.1 A Schematic of a GRN 4
1.2 Skeletogenesis in the sea urchin embryo 7
1.3 The endoskeleton of the sea urchin pluteus 7
1.4 Initial activation of the PMC GRN 8
1.5 Activation and stabilization of early and late regulatory genes 11
1.6 Morphogenetic events that occur during skeletogenesis 12
1.7 Alx1 and Ets1 regulatory inputs into PMC effector genes 13
1.8 Regulatory inputs into PMC effector genes 14
1.9 Deuterostome phylogeny, with an expanded view of the echinodermata phylum 15

2.1 Linear scatter plot of FPKM from RNA-seq of isolated PMCs and other (non-PMC) cells 31
2.2 WMISH analysis of mRNAs differentially expressed in PMCs 33
2.3 Hierarchical clustering of temporal expression patterns of genes differentially expressed in PMCs 35
2.4 Distribution of DE genes by functional class 36
2.5 Venn diagram showing overlapping distributions of genes affected by Ets1 knockdown, Alx1 knockdown, or U0126 treatment among the 420 genes differentially expressed by PMCs 39
2.6 Distinct temporal gene expression profiles of Ets1/Alx1 co-regulated targets and non-target genes in the DE set 40

3.1 DNase-seq Sample Preparation and Sequence Analysis 66
3.2 ATAC-seq Sample Preparation and Sequence Analysis 69
3.3 Correspondence Between DNase-seq and ATAC-seq Datasets 70
3.4 Previously Studied PMC-specific Cis-regulatory Modules 73
3.5 Validation of Putative PMC CRMs Using GFP Reporter Gene Assays 73

List of Tables

3.1 Sequence analysis information for DNase-seq samples 63
3.2 Sequence analysis information for ATAC-seq samples 67
3.3 Correspondence between ATAC-seq and DNase-seq datasets 70
3.4 PMC CRMs validated by reporter gene assays 74
3.5 Enrichment of PMC Transcription Factor Consensus Binding Sites in Differential Peaks 75
3.6 Predicted TF Binding Sites in Validated PMC CRMs 76

List of Supplementary Figures

2.1 Transcript abundances per PMC for genes differentially expressed in PMCs 49
2.2 Reproducibility of gene expression profling of Ets1 and Alx1 morphants . . 50

3.1 Correlation of DNase-seq and ATAC-seq peaks within replicates 86
3.2 Functional category (GO) enrichment for differential peak sets 88
3.3 Examples of overlapping differential peaks accessible at the 128-cell stage . 88
3.4 Sequences enriched in overlapping differential peaks, as identified by de novo motif discovery 89

Abstract

As development proceeds, cells acquire specialized properties and functions that are critical for the formation of a complex multicellular organism. Despite having the same genome, groups of embryonic cells perform varied developmental functions due to the precise regulation of gene expression that enables specialized genes to be expressed in the right place at the right time. The expression of these specialized genes drives morphogenesis, and enables the formation of complex tissues and body parts. A key question in developmental biology is: how do cells decode instructions from the genome to carry out morphogenetic processes? Gene regulatory networks (GRNs) are powerful tools for elucidating the genomic control of morphology. GRNs depict interactions between regulatory genes such as transcription factors and signaling molecules and effector genes that carry out morphogenetic functions.

The GRN underlying the skeletogenic lineage in the sea urchin embryo has emerged as a model network to study how the genome directs the specification of a cell lineage during development. The morphological process of skeletogenesis has been studied extensively in the sea urchin embryo, and several lineage-specific regulatory genes have been identified and linkages among them have been elucidated. The initial activation of this network specifically in the skeletogenic lineage has been dissected, and most functional regulatory linkages among TFs have been elucidated. Some functional regulatory linkages between skeletogenic regulators and effectors have been mapped. I have identified a handful of novel regulatory genes and hundreds of novel effector genes in the skeletogenic lineage in a high-throughput manner, resulting in a much more comprehensive view of the regulatory and effector genes involved in skeletogenesis. We also uncovered functional interactions between two TFs and a set of over 200 effector genes, greatly expanding the number of regulatory connections between TFs and effector genes in the network.

The large majority of regulatory connections in the network have been uncovered by perturbing the function of regulatory genes and assaying the effect of this perturbation on other genes. Direct regulatory connections cannot be differentiated from indirect regulatory connections using this method. Only a handful of direct interactions between skeletogenic regulators and effectors have been mapped by conventional *cis*-regulatory analysis. I have been able to identify over 3,000 putative *cis*-regulatory modules (CRMs) mediating skeletogenic gene expression using genome-wide techniques. I have inferred some regulatory connections into these CRMs and demonstrated the value of using differential chromatin accessibility to identify cell-type-specific CRMs in a high-throughput manner in early embryos.

This thesis work has greatly expanded the number of skeletogenic effector genes in the network and enabled the identification of thousands of regulatory connections between

upstream TFs and downstream effector genes. This effort to construct a detailed and comprehensive skeletogenic GRN will enable a detailed understanding of how instructions from the genome are decoded during the establishment of a cell lineage during development. Several discrete GRN subcircuits elucidated in the skeletogenic GRN can be dissected in greater detail and used to understand the fundamental principles of GRN architecture. This detailed GRN can be used to obtain a deeper understanding of the evolutionary mechanisms that enable the acquisition of novel morphological structures during speciation. The network includes biomineralization genes that are conserved across vertebrates, and further dissection of the regulation of these genes will aid in the discovery of a common biomineralization toolkit likely used by diverse animal lineages.

Chapter 1

Introduction

Development is the process by which a single fertilized egg becomes a complex multicellular organism consisting of multiple cell types, tissues and body parts. The genome directs developmental processes such as cell specification, morphogenesis, differentiation and growth during embryogenesis. Every cell in a developing embryo contains the same genome, and yet performs varied and complex functions throughout embryogenesis. As development progresses, morphological complexity increases, followed by the progressive emergence of defined tissues and body parts.

A fundamental question in developmental biology is: how are instructions from the genome decoded by cells of the developing embryo? In order for development to proceed, it is critical that different sets of embryonic cells acquire the ability to perform unique and varied developmental functions. Differential gene expression is the primary mechanism that drives cell fate specification, conferring unique and critical functions to groups of embryonic cells. During early development, precise transcriptional regulation enables genes to be expressed in the appropriate cell types at the correct developmental stage.

1.1 An Introduction to Gene Regulatory Networks (GRNs)

Gene Regulatory Networks (GRNs) have emerged as a valuable tool for studying the transcriptional control of development (Davidson, 2001; Levine and Davidson, 2005; Ettensohn, 2009, 2013; Peter and Davidson, 2015). GRNs depict interactions between developmental genes during embryogenesis, i.e., interactions between regulatory genes (encoding transcription factors), signaling molecules and differentiation and morphogenesis genes. In a GRN, linkages between regulatory genes and downstream "effector" genes that perform developmental functions depict regulatory connections that drive accurate spatiotemporal gene expression in the embryo. For a hypothetical example of a GRN, see Figure 1.1.

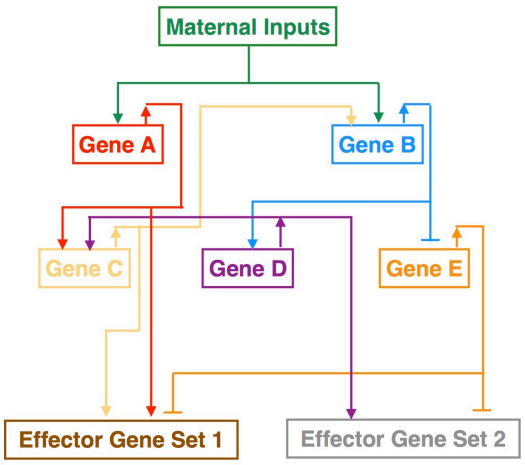


Figure 1.1: A Schematic of a GRN. Localized maternal inputs activate a set of regulatory genes, which interact with other regulatory genes and set up the "regulatory state" that specifies a particular cell type during development. These regulatory genes mediate the spatio-temporal activation of effector genes that carry out cell-type-specific developmental functions. Activation inputs are depicted as arrows and repressive inputs are depicted by bars.

Several different tools and techniques can be used to obtain the information required to construct a GRN. First, the relevant genes constituting the network must be identified. A genome-wide screen can be performed using high-throughput methods such as RNA-seq (Mortazavi et al., 2008) to identify genes that are expressed at a particular developmental stage or even in a particular cell type. Smaller-scale screens can be performed using qPCR or NanoString (Malkov et al., 2009) to identify sets of genes expressed at particular times during development. Whole-mount in situ hybridization (WMISH) screens can be used to identify genes expressed in the relevant cell populations during different developmental stages. Computational methods can be used to identify sets of genes likely to participate in a network based on information from other organisms.

Second, functional interactions among the genes expressed in the cell lineage of interest across development must be identified. Perturbation analyses, in which a gene is knocked down and the effects on the expression of all other genes is studied, are commonly used to identify functional linkages. Molecular techniques commonly used to perturb gene function include the generation of transgenic embryos by various techniques including homologous recombination, microinjection of DNA constructs, and CRISPR/Cas9 (Cong et al., 2013; Mali et al., 2013) mediated genome editing. Morpholinos (Summerton and Weller, 1997), and RNAi (Fire et al., 1998) methods can be used to perturb gene function in organisms not amenable to genetic manipulation.

Third, *Cis*-regulatory modules (CRMs) mediating the expression of genes in the net-

work must be identified. High-thoughput chromatin-profiling methods such as ATAC-seq (Buenrostro et al., 2015), DNase-seq (Crawford et al., 2006) and ChIP-seq (Johnson et al., 2007) can be used to identify thousands of putative CRMs. Classic *cis*-regulatory analysis using reporter assays identify CRMs on a smaller scale. Fourth, direct regulatory connections must be established between upstream regulators (TFs and signaling molecules) and the CRMs of effector genes (morphogenesis and differentiation genes). TF binding sites can be identified in a high-throughput manner using techniques such as ChIP-seq, genome-wide footprinting methods (Hesselberth et al., 2009), and protein-binding microarrays (PBMs) (Berger and Bulyk, 2009). Careful mutational analysis of CRMs can also identify specific TF binding sites. In conjunction with these methods, previously identified functional linkages can aid the establishment of direct TF-effector gene linkages.

1.2 Using the Sea Urchin Embryo to Study GRNs

The sea urchin embryo serves as an excellent model organism to study the genomic control of development using GRNs. Sea urchin embryogenesis is relatively simple, with few regulatory steps between the initial specification of cells to the activation of terminal differentiation genes (Davidson et al., 2002). The sea urchin embryo develops into a simply constructed larva that consists of single-cell thick structures and only 10-12 cell types: it is much easier to study than a morphologically complex juvenile version of the adult body plan (Davidson et al., 2002). The developmental biology of the sea urchin has been well-studied: detailed cell fate maps exist for all embryonic cell types. Some morphogenetic processes, such as the formation of the embryonic skeleton, have been dissected in detail (Wilt and Ettensohn, 2007; Ettensohn, 2009).

The sea urchin embryo is robust and can tolerate microinjection as well as various drug treatments, making it easy to perturb gene function and study its effects on the network. While conventional genetic approaches cannot be used to knockout genes, morpholinos injected into the sea urchin egg are often effective for gene knockdowns, and the CRISPR/Cas9 technique has been recently applied successfully to disrupt gene function in sea urchin embryos (Lin and Su, 2016; Oulhen and Wessel, 2016). Procedures such as cell transplantations, depletions and isolations can be performed to study signaling interactions and specific cell types in isolation. It is also optically clear, enabling the use of different microscopy techniques, and aiding the study of spatio-temporal gene expression using whole-mount in situ hybridization (WMISH) techniques.

The sea urchin genome has been sequenced, and a high-quality assembly is available along with computational and manual gene annotations (Sea Urchin Genome Sequencing Consortium et al., 2006). Genomes of related echinoderms have also been sequenced, serving as a valuable resource for comparative studies. A comprehensive transcriptome with temporal expression profiles of nearly all genes expressed across development is

available (Tu et al., 2012, 2014). Genome-wide chromatin profiling using ATAC-seq on whole embryos at different developmental stages is available on echinobase.org and cell-type-specific chromatin profiling has also been conducted for the skeletogenic cells of the embryo (Shashikant et al., *in review*), enabling the identification of thousands of putative cell-type specific *cis*-regulatory modules (CRMs). The spatio-temporal expression of most regulatory genes expressed during the development of the endomesoderm is known (Peter and Davidson, 2011). A large collection of BAC libraries containing genes as well as surrounding regulatory regions is available (Cameron et al., 2000; Sea Urchin Genome Sequencing Consortium et al., 2006), and a GFP reporter plasmid for use in the sea urchin embryo has been created (Cameron et al., 2004), enabling the testing of large and short putative CRMs.

The depth of knowledge of sea urchin development and the vast molecular resources available have greatly aided the construction of the sea urchin endomesoderm GRN, which is among the most comprehensive and detailed GRNs in any animal (Peter and Davidson, 2015). Within this network, the GRN specifying the embryonic skeleton is especially well-studied (Oliveri et al., 2008; Ettensohn, 2009; Rafiq et al., 2012; Ettensohn, 2013; Rafiq et al., 2014).

1.3 Skeletogenesis in the Sea Urchin Embryo

The formation of the embryonic endoskeleton in sea urchins is a powerful experimental model for understanding how the genome encodes morphogenesis during development. The endoskeleton is an elaborate calcareous structure that shapes the sea urchin larva, enables feeding and swimming and probably defends against predation (Ettensohn, 2013).

The skeletogenic lineage originates at the 32-cell stage, in the four large micromeres formed at the vegetal pole of the embryo. The large micromeres undergo further divisions and become incorporated into the epithelial wall of the blastula. At the mesenchyme blastula stage, the large micromere descendants undergo an epithelial-to-mesenchymal transition and ingress into the blastocoel. From this point on, they are referred to as "primary mesencyme cells" (PMCs) (Wilt and Ettensohn, 2007).

During gastrulation, the PMCs migrate and fuse using filopodia, and form a syncytial subequatorial ring along the blastocoelar wall. Two PMC clusters are formed along the ventro-lateral regions of the ring, and the secretion of the endoskeleton begins here. During the mid-gastrula stage, the PMCs at the ventro-lateral clusters secrete biominerals that create a tri-radiate spicule at each cluster. The three arms are then extended by further deposition of biominerals. The skeleton elongates and branches, to form the stereotypical structure seen in the pluteus larva. See Figure 1.3 for images of a pluteus larva. The biorefringent skeleton is clearly visible under plane polarized light.

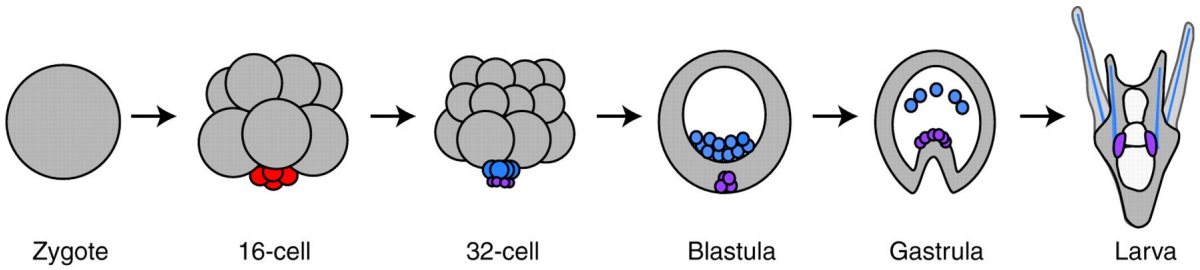


Figure 1.2: Skeletogenesis in the sea urchin embryo. (Figure from Juliano et al. (2010))

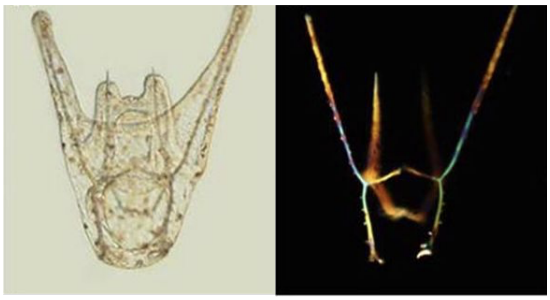


Figure 1.3: The endoskeleton of the sea urchin pluteus. (Figure from Adomako-Ankomah and Ettensohn (2013))

The various morphogenetic steps involved in skeletogenesis have been studied in great detail (Ettensohn, 2013). The full set of regulatory genes expressed in the skeletogenic lineage is known (Oliveri et al., 2008; Rafiq et al., 2014) and their spatio-temporal expression during skeletal morphogenesis has been determined at a high-resolution (Peter and Davidson, 2011). Putative CRMs regulating skeletogenic gene expression have been recently identified genome-wide (Shashikant et al., *in review*). These advances have enabled the construction of a relatively detailed GRN that describes how regulatory genes set up the skeletogenic lineage and mediate the expression of skeletal morphoeffector genes during skeletogenesis.

1.4 The Primary Mesenchyme Cell GRN

1.4.1 Initial Activation of the Network in the Skeletogenic Lineage

The initial deployment of the PMC GRN is dependent on maternal inputs. Disheveled (Dsh) is localized at the vegetal cortex of the egg during oogenesis. Dsh prevents the degradation of cytoplasmic β-catenin and enables the nuclearization of β-catenin in the micromeres formed at the 16-cell stage (fourth cleavage) (Wikramanayake et al., 1998; Logan et al., 1999; Weitzel et al., 2004; Ettensohn, 2006). At this stage, only the micromeres contain nuclear β-catenin, setting the stage for the specification of the skeletogenic cell

lineage. β-catenin then interacts with the TCF transcription factor and maternal Otx (Chuang et al., 1996; Klein and Li, 1999) and activates the zygotic transcription of *pmar1* (Kitamura et al., 2002; Oliveri et al., 2002), a paired class homeodomain-containing protein, specifically in the micromeres at the end of the fourth cleavage.

β-catenin then interacts with TCF and maternal Blimp1 to activate *Wnt8* gene expression in the micromeres during the fifth cleavage stage (Smith et al., 2007; Minokawa et al., 2005). Wnt8 further drives β-catenin nuclearization in micromeres. Maternal Otx, also nuclearized in micromeres, acts with β-catenin/TCF to drive the zygotic expression of *blimp1* during the sixth cleavage (Smith et al., 2007).

The activation of *pmar1* is the first step towards, and is necessary and sufficient for, the specification of the skeletogenic lineage (Oliveri et al., 2002, 2003). Since it is a repressor, it does not directly activate the expression of skeletogenic lineage genes. It is thought to act via the repression of *hesC*, a member of the HES (Hairy-Enhancer-of-Split) family (Revilla-i Domingo et al., 2007). HesC is a repressor, present ubiquitously in the embryo during early cleavage. Evidence for the repression of *hesC* by Pmar1 is the observation that overexpression of Pmar1 results in a decrease in *hesC* expression levels (Revilla-i Domingo et al., 2007). HesC represses five key regulators of skeletogenic genes: *alx1* (Ettensohn et al., 2003), *ets1* (Kurokawa et al., 1999), *tbr* (Fuchikami et al., 2002), *tel* and signaling molecule *delta* (Sweet et al., 2002; Oliveri et al., 2002). Figure 1.4 depicts the initial activation of the PMC GRN.

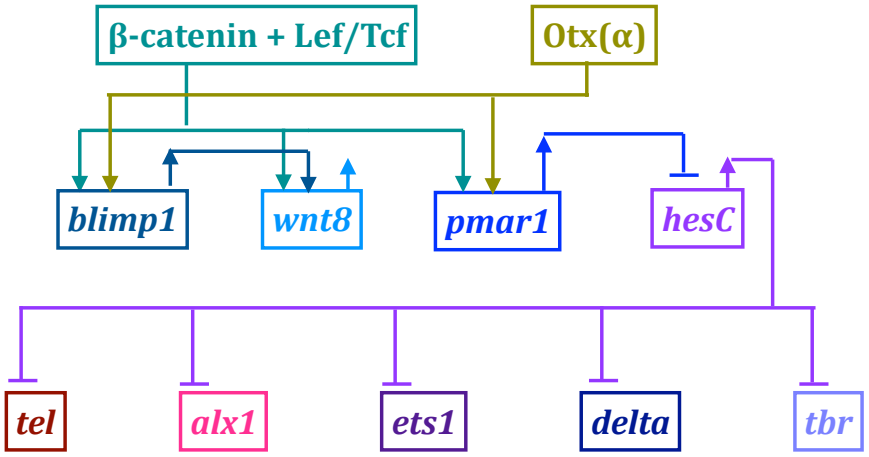


Figure 1.4: Initial activation of the PMC GRN. Maternal inputs are localized at the vegetal pole, and activate a set of early TFs. Pmar1 represses HesC in the micromeres, and activates the skeletogenic network specifically in this lineage. Regulatory connections depicted are based on data cited in the text and Oliveri et al. (2008). Arrows depict positive regulatory inputs and bars depict repressive inputs.

The observation that the knockdown of HesC causes ectopic expression of *delta* and increased levels of *alx1, ets1* and *tbr* throughout the embryo is evidence of the repressive input of HesC into these skeletogenic genes. *Cis*-regulatory analysis of *alx1* (Damle and Davidson, 2011), *tbr* (Wahl et al., 2009) and *delta* (Revilla-i Domingo et al., 2004; Smith, 2008) has identified functional binding sites for HesC that are required for the repression of these genes in cells other than the skeletogenic lineage, thereby establishing a direct repressive linkage between HesC and these skeletogenic genes. This evidence points to a double-negative gate being the primary mechanism of the activation of skeletogenic genes specifically in the micromeres: when the expression of *pmar1* is activated in the micromeres, *hesC* is repressed, thereby causing the activation of skeletogenic genes previously repressed by HesC in the micromeres. No skeletogenic gene expression is activated in the small micromeres, despite the presence of Pmar1, possibly because the small micromeres remain transcriptionally quiescent until later in development, when they produce primordial germ cells (Yajima and Wessel, 2011, 2012)

Another set of observations challenge the initial repression of *alx1* and *delta* by HesC: it was shown by WMISH that the activation of *alx1* and *delta* expression selectively in the large micromere territory occurs prior to the clearing of *hesC* mRNA from these cells (Sharma and Ettensohn, 2010). The expression level of *HesC* mRNA was demonstrated to be equivalent throughout the embryo until the blastula stage, several developmental stages after the specification of the skeletogenic lineage by Alx1. It is possible that the early repression of *alx1* and *delta* in non-skeletogenic lineage cells is mediated by a different repressor, or that the restriction of the skeletogenic lineage to the micromeres occurs through a different mechanism. Observations in closely related cidaroid urchins (*Eucidaris tribuloides* and *Prionocidaris baculosa*) and brittle star (*Amphiura filiformis*) point to this possibility. Skeletogenic cell specification occurs independently of the double-negative gate in the cidaroid urchin (Yamazaki et al., 2014; Erkenbrack and Davidson, 2015) and the *pmar1* ortholog does not repress *hesC*, and neither does HesC repress the *delta, ets* and *tbr* orthologs in the non-skeletogenic lineage of *A. filiformis* (Dylus et al., 2016).

1.4.2 Activation of Early and Late Regulatory Genes

Pmar1 activates the expression of an early set of transcription factors: *alx1, ets1, tbr* and *tel*. *Alx1* (aristaless-like homeobox 1) is the earliest regulatory gene activated in the skeletogenic cell lineage. It is a member of the Cart1/Alx3/Alx4 subfamily of paired class homeodomain proteins. It is restricted to the skeletogenic lineage throughout development. Ets1 (E26 transformation specific 1) is a transcription factor that contains the highly conserved ETS domain. *Ets1* mRNA and protein are present maternally. Zygotic expression of *ets1* occurs during late cleavage, and is restricted to the skeletogenic lineage until the late mesenchyme blastula stage. Both *alx1* and *ets1* are required for PMC ingression and subsequent morphogenesis. When *alx1* expression is knocked down, PMCs do not ingress and no skeleton is formed (Ettensohn et al., 2003). Overexpression of *alx1* causes

a dose-dependent increase in the number of PMCs, but has the opposite effect at high concentrations (Ettensohn et al., 2003; Damle and Davidson, 2011). *Alx1* functions as an autoregulator of its own expression: increasing expression at low mRNA levels and decreasing expression at high mRNA levels (Ettensohn et al., 2003; Damle and Davidson, 2011). Overexpression of *ets1* transforms most cells of the embryo into mesenchymal cells.

The early regulatory TFs Alx1, Ets1 and Tbr activate a set of late regulatory genes: *erg, dri, hex, tgif, foxB, foxO* and *foxN2/3*. *FoxB* and *dri* are activated by Alx1. *Hex, dri, erg, tgif, foxB* and *foxN2/3* are activated by Ets1. *Erg, foxB* and *foxN2/3* are activated by Tbr. An RNA-seq study (Rafiq et al., 2014) identified four additional skeletogenic lineage TFs co-regulated by Alx1 and Ets1: *mitf, nk7, cebpa* and *alx4*. These TFs were downregulated when *ets1* was knocked down, as well as when *alx1* was knocked down. There are three possible mechanisms that explain the co-regulation of genes by Ets1 and Alx1: First, since Ets1 activates *alx1* expression after cleavage, Ets1 may be regulating other TFs via Alx1. Second, Ets1 and Alx1 may be co-regulating genes via a feed-forward mechanism: Ets1 activates *alx1* expression, and Ets1 and Alx1 together activate a downstream gene. Third, Ets1 and Alx1 may activate the expression of an intermediate TF, which then activates downstream gene expression. See Rafiq et al. (2014) for an explanation of the evidence supporting all three mechanisms.

Two TFs, *mef2* and *smad2/3* were found to be enriched in PMCs, but not regulated by Alx1 or Ets1 (Rafiq et al., 2014). Two other TFs, Snail and Twist were found to be essential for PMC ingression and fusion in *Lytechinus variegatus* embryos. They are regulated by Alx1, but have not been studied in *S. purpuratus* embryos and hence have not been added to the *S. purpuratus* PMC GRN discussed here. Direct regulatory connections into and among these TFs newly added to the network have not yet been identified, but they likely interact with other regulatory genes and mediate the expression of sets of skeletal morphogenesis genes.

This set of early and late regulatory genes is engaged in dynamic interactions that establish the regulatory state of the skeletogenic lineage, primarily via positive feedback loops that serve to enhance and maintain gene expression. Erg upregulates *hex, tgif* and *foxO* expression, Hex upregulates *erg* and *tgif* expression while Tgif upregulates *hex, foxO* and *alx1* expression. These positive feedback loops convert the transient expression of *pmar1* into a stable regulatory state in the skeletogenic lineage and buffer against initial variation in the level of expression of these genes. Since the genes involved in these positive feedback loops are co-dependent, loss of expression of even one gene severely impacts the expression of all genes involved in the loop, and this has a catastrophic effect on the expression of downstream genes. This imposes an evolutionary constraint on the network, making rewiring very difficult without losing function entirely (Peter and Davidson, 2015).

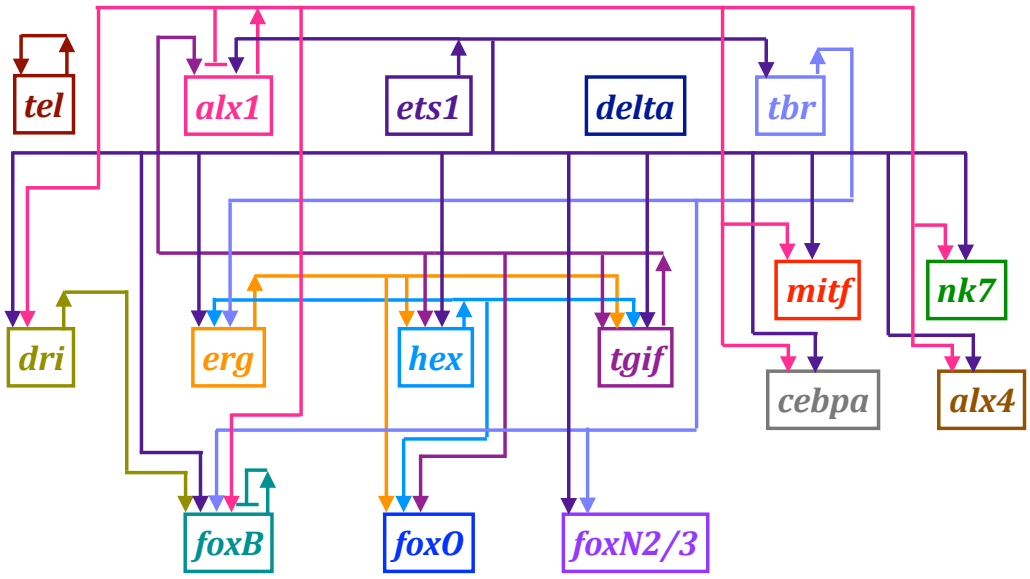


Figure 1.5: Activation and stabilization of early and late regulatory genes. Early regulatory genes activate a set of late regulatory genes, and positive feedback interactions among these genes stabilize the skeletogenic lineage regulatory state. Regulatory linkages obtained from Oliveri et al. (2008). Arrows depict positive inputs and bars represent repressive inputs.

1.4.3 Activation of Skeletal Effector Genes

Once the skeletogenic regulatory state is established, skeletal morphogenesis and differentiation genes, called "effector genes", are activated in the skeletogenic lineage. A majority of genes are activated by the time of PMC ingression. Around 400 effector genes have been identified in two recent RNA-seq studies (Rafiq et al., 2014; Barsi et al., 2014). Alx1 and Ets1 upregulate ∼50% of these genes (Rafiq et al., 2014). So far, direct inputs from upstream regulators have been identified using *cis*-regulatory analysis for five effector genes: *sm50* (Makabe et al., 1995), *sm30* (Akasaka et al., 1994; Yamasu and Wilt, 1999), *tbr* (Wahl et al., 2009), *alx1* (Damle and Davidson, 2011) and *cyp1* (Amore and Davidson, 2006).

PMCs undergo a series of morphogenetic events while establishing the embryonic skeleton. An epithelial-to-mesenchymal transition occurs during the ingression of PMCs into the blastocoel at the mesenchyme blastula stage. This morphogenetic process is mediated by Ets1 (Kurokawa et al., 1999), which is activated by MAPK signaling (Röttinger et al., 2004). In *L. variegatus* embryos, *snail*, *twist* and *foxN2/3* were found to be required for ingression of PMCs (Wu and McClay, 2007; Wu et al., 2008; Rho and McClay, 2011). After PMCs ingress, they migrate directionally within the blastocoel. VEGF and FGF ligands mediate the migration of PMCs, which specifically express the cognate receptors

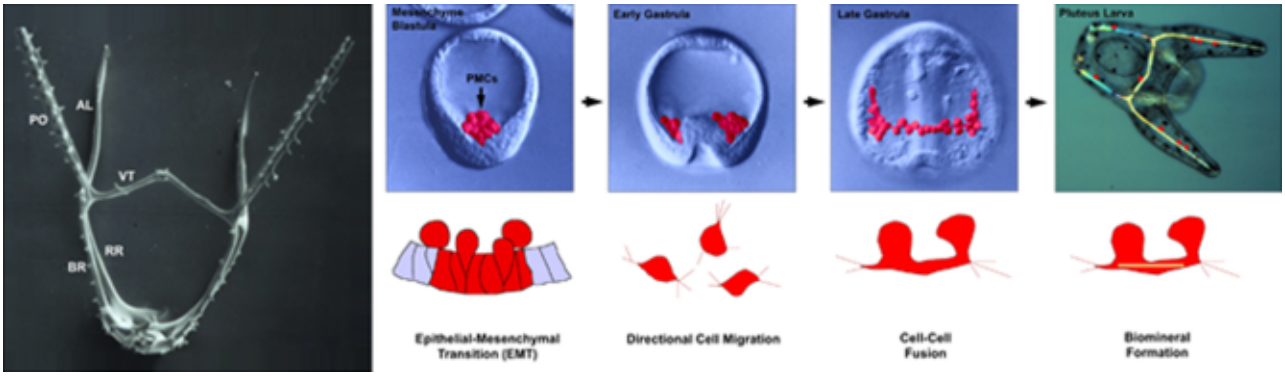


Figure 1.6: Morphogenetic events that occur during skeletogenesis (Figure from Ettensohn (2013))

vegfr-Ig-10 and *fgfr2* (Duloquin et al., 2007; Röttinger et al., 2008; Adomako-Ankomah and Ettensohn, 2013). The PMC GRN depicts four regulatory inputs into *vegfr-Ig-10*: *dri, hex, ets1* and *alx1*.

During their migration within the blastocoel, the PMCs undergo filopodia-mediated fusion and form a syncytium. KirrelL, a member of the Ig-domain superfamily of cell adhesion proteins, was recently shown to be essential for PMC fusion (Ettensohn and Dey, 2017). *KirrelL* is expressed specifically in the PMCs and is regulated by Ets1 as well as Alx1 (Rafiq et al., 2014). A recent study identified a CRM mediating KirrelL expression in PMCs (Shashikant et al., *in review*). *Twist* and *foxN2/3* was also determined to be essential for PMC fusion in *L. variegatus* embryos (Wu et al., 2008; Rho and McClay, 2011). After the syncytium is formed, biominerals are secreted and deposited within the syncytium, in a process called "biomineralization". Several types of proteins are involved in biomineralization, and the majority of the genes encoding these proteins are regulated by Alx1 or Ets1 or both (Rafiq et al., 2014). See Figure 1.7 for PMC effector genes regulated by Alx1 and Ets1.

The morphogenetic process of biomineralization involves several steps, and each step requires the deployment of a set of effector genes that perform specialized functions. Some examples are briefly described. A set of spicule matrix genes encode proteins that are occluded in small amounts (<0.1%) within the calcite skeletal rods, and they regulate the formation of calcite crystals (Livingston et al., 2006; Mann et al., 2010; Ettensohn, 2013). The MSP130 family of cell-surface glycoproteins possibly regulate calcium acquisition by interacting with channels or transporters (Wilt and Ettensohn, 2007). Type I transmembrane proteins including the P16 family, FcgbpL (p58a) and Hypp_302 (p58b) are involved in biomineral deposition (Cheers and Ettensohn, 2005; Adomako-Ankomah and Ettensohn, 2011). The TgfbrtII receptor was recently found to be required for biomineral deposition (Sun and Ettensohn, 2017). It is expressed specifically in PMCs and is regulated by both Alx1 and Ets1 (Rafiq et al., 2014). Figure 1.8 depicts functional regulatory connections into effector genes.

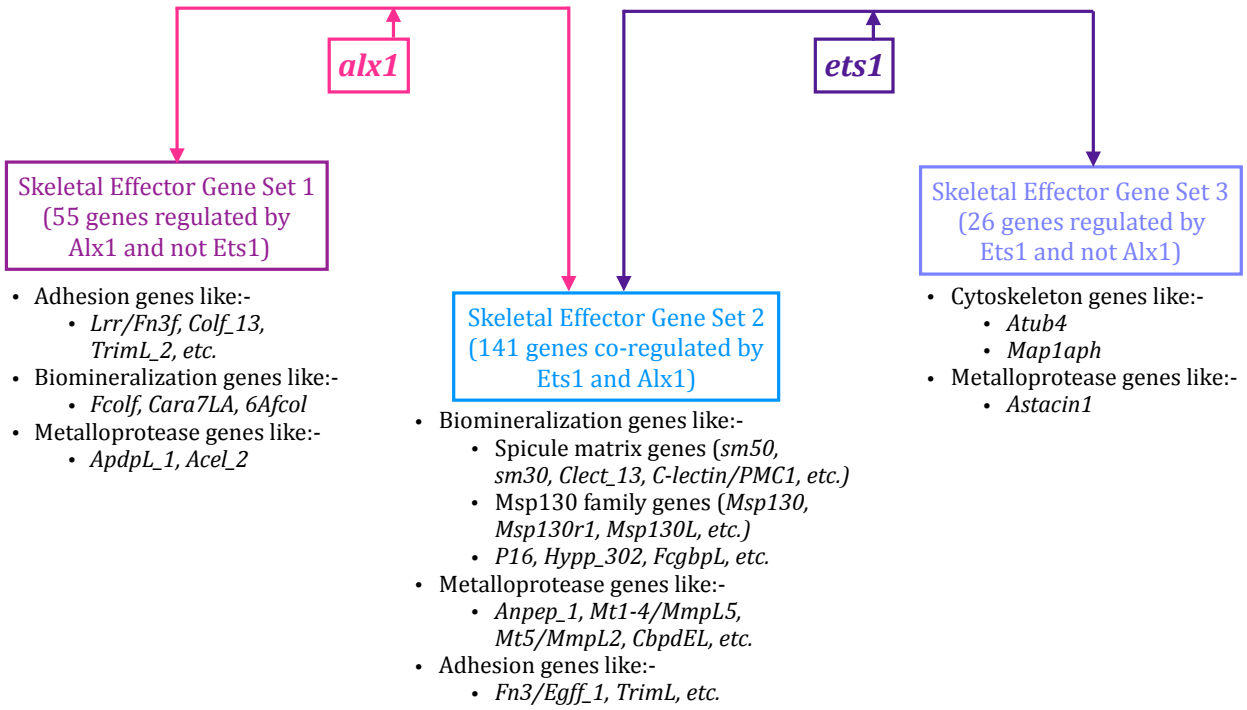


Figure 1.7: Alx1 and Ets1 regulatory inputs into PMC effector genes. 222 PMC-enriched transcripts are positively regulated by Alx1, Ets1 or both (Rafiq et al., 2014). Genes regulated by Alx1 and not Ets1 (Effector Gene Set 1), genes co-regulated by Alx1 and Ets1 (Effector Gene Set 2) and genes regulated by Ets1 and not Alx1 (Effector Gene Set 3) are depicted, with a few specific examples.

While we have a reasonably complete list of effector genes likely to play roles in skeletal morphogenesis, we have only recently been able to identify a large set of putative CRMs mediating the expression of these genes using high-throughput methods (Shashikant et al., *in review*). This study has resulted in a large number of putative PMC CRMs that can be dissected further to establish direct regulatory connections to upstream TFs.

1.5 Using the PMC GRN to Study the Evolution of Developmental Programs

Echinoderms serve as excellent model organisms to study the evolution of developmental programs, for several reasons. First, the existence of an extensive fossil record, coupled with detailed molecular phylogenies and molecular clock analyses has led to a relatively deep understand of the evolution of echinoderms. Second, due to the ease of obtaining, culturing and studying embryos in large numbers, the varied developmental mechanisms

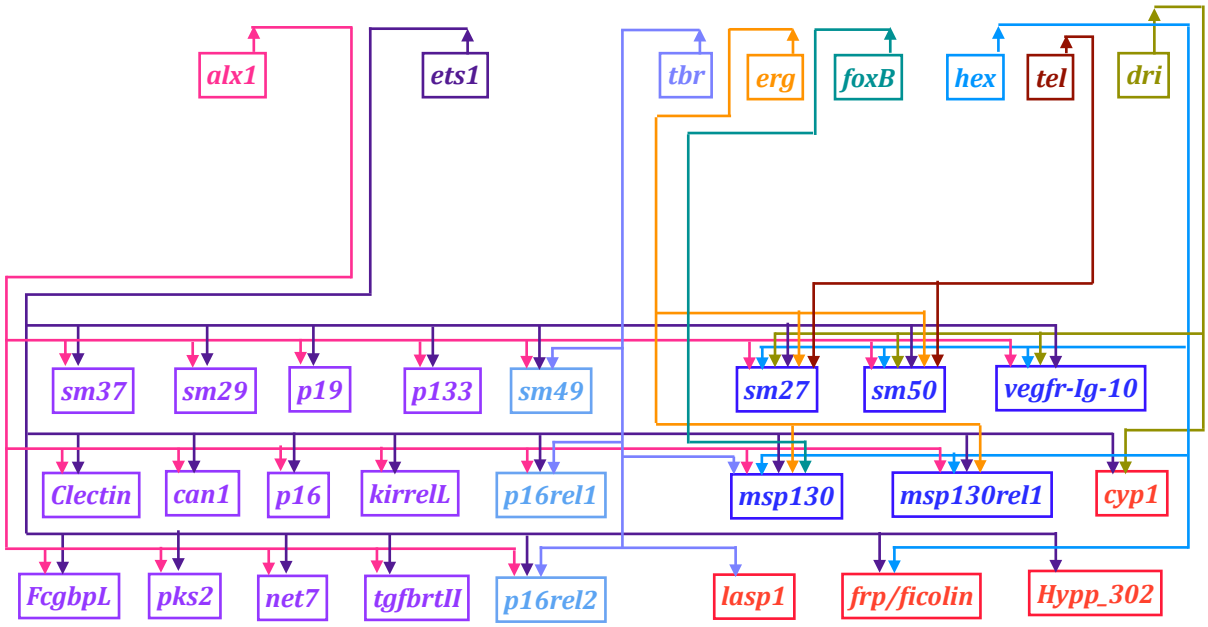


Figure 1.8: Regulatory inputs into PMC effector genes. Nearly half the PMC effector genes have Alx1 and Ets1 regulatory inputs (see Figure 1.7 and Rafiq et al. (2014)). Positive regulatory inputs from other early and late TFs into effector genes are depicted. Regulatory connections obtained from Rafiq et al. (2012) and data cited in text.

and embryology of echinoderms is very well understood. Third, a relatively detailed and comprehensive GRN has been delineated for the sea urchin, that serves as a model for comparative studies of the evolution of GRNs across echinoderms. Fourth, recent genomic advances have shed further light on the molecular basis of the evolution of GRNs in echinoderms (reviewed in (Cary and Hinman, 2017)). More specifically, the acquisition of the larval skeleton in sea urchins has been used to understand how complex morphologies have evolved by the acquisition of novel structures. As described previously in detail, the development of the larval skeleton is well-studied and the GRN underlying skeletogenesis is one of the most comprehensive available.

Deuterostomes are classified into three phlya: Chordata, Hemichordata and Echinodermata. Hemichordates and Echinoderms, collectively referred to as Ambulacaria, diverged from each other around 570 million years ago (MYA) (Erwin et al., 2011). Echinoderms are classified into 5 major classes: crinoids (sea lillies and feather stars), ophiuroids (brittle stars), asteroids (sea stars), holothuroids (sea cucumbers) and echinoids (sea urchins and sand dollars).

Ophiuroids and asteroides are grouped into the Asterozoa clade and holothuroids, cidaroids and euechinoids are grouped into the Echinozoa clade, based on recent genomic evidence (Telford et al., 2014; Reich et al., 2015). Echinoids are further divided into euechinoids and cidaroids. Euechinoids constitute the vast majority of present-day sea urchins (in-

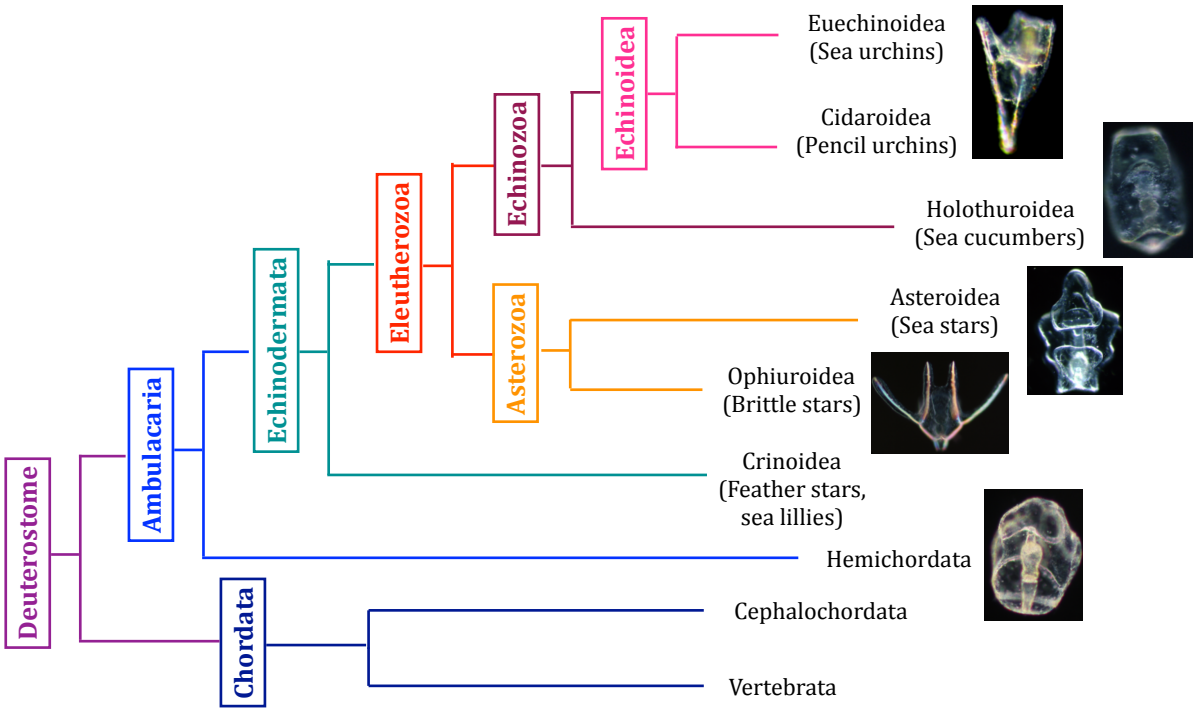


Figure 1.9: Deuterostome phylogeny, with an expanded view of the echinodermata phylum. Representative larvae from some echinoderm classes are shown. Branch lengths are not to scale. (Images of larvae from Koga et al. (2014))

cluding the model species *Strongylocentrotus purpuratus* and *Lytechinus variegatus*) and sand dollars. Cidaroids are pencil urchins: *Eucidaris tribolidea* is the best-studied model species.

Most echinoderms undergo indirect development, in which the larval form does not anatomically resemble the adult. All adult echinoderms posses a biomineralized endoskeleton, but only euechinoid sea urchins and ophiuroid brittle stars have an extensive embryonic skeleton: however, unlike sea urchins, brittle stars do not form micromeres and PMCs. Holothuroids and cidaroids create a relatively rudimentary embryonic skeleton, but asteroids and crinoids do not possess a larval endoskeleton.

1.5.1 Skeletogensis in Cidaroids

Recent studies in cidaroids (Yamazaki et al., 2014; Erkenbrack and Davidson, 2015) reveal some key differences in the formation of the embryonic skeleton compared to euechinoids. Cidaroid embryos form variable numbers of micromeres and micromere descendants do not ingress into the blastocoel at the onset of gastrulation. Instead, skeletogenic cells, along with other mesenchymal cells, delaminate from the tip of the archen-

teron during mid-gastrula. They migrate and construct a skeleton much later in development.

Several differences in the specification of skeletogenic cells in *Eucidaris tribolidea* compared to euechinoids have been identified (Erkenbrack and Davidson, 2015). The *ets1* is zygotically expressed early in development, soon after micromere specification. However, it is no longer expressed in the skeletogenic mesenchyme soon after the blastula stage: it is only expressed in the non-skeletogenic mesenchyme, and does not have a role in skeletogenic effector gene expression during and after gastrulation. No expression of the *alx1* ortholog is seen in micromeres or its immediate descendants, but it is expressed in a skeletogenic cell-specfic manner later in development and is critical for skeletogenesis. The *delta* ortholog is expressed early in micromeres, even before *alx1* activation, and remains specific to skeletogenic cells until much later in development. The *hesC* ortholog is expressed in the micromeres along with *delta* and *ets1* and does not act as a repressor of skeletogenic genes, except for its repression of *alx1* expression in the non-skeletogenic mesenchyme, where it is expressed during the blastula stage. It is also not expressed ubiquitously in the embryo at any stage. Expression of the *tbr* ortholog is initially activated in the micromeres, but is soon expressed in the non-skeletogenic mesoderm and does not seem to have a role in skeletogenesis.

Given these varied expression patterns and functions of key *S. purpuratus* PMC GRN components in *E. tribolidea*, it is not surprising that there are some major differences in the skeletogenic GRNs of these closely related echinoids. First, the initial maternal combinatorial input of Otxα and Tcf/β-catenin does not function in skeletogenic micromere specification. Second, the activation of the skeletogenic program in the micromeres does not occur through the *pmar1* – *hesC* double-negative gate. No *pmar1* ortholog has been identified, and *hesC* was found not to have a repressive effect on skeletogenesis genes. Third, *alx1* does not mediate early skeletogenic specification in the micromeres, but is still required for skeletogenesis, unlike *ets1* and *tbr*, which are expressed early in micromeres but are not required for skeletogenesis.

1.5.2 Skeletogenesis in Holothuroids

A study carried out by McCauley et al. (2012) shed some light on the regulatory genes involved in skeletogenesis in the sea cucumber embryo. The sea cucumber embryo does not form micromeres during early cleavage. Mesenchyme cells ingress from the vegetal plate at the onset of gastrulation, migrate during the mid-gastrula stage and construct a relatively morphologically simple larval skeleton consisting of small spicules. The *Parastichopus parvimensis* (sea cucumber) orthologs of *ets1, erg, foxN2/3, tbr* and *tgif* are expressed in presumptive mesodermal cells. The *alx1* ortholog is expressed in presumptive skeletogenic cells and is required for skeletogenesis. However, it is not restricted to skeletogenic cells, and is also expressed in a broader mesoderm territory. The *Tbr* ortholog is not ex-

pressed in the skeletogenic lineage and doesn't seem to be required for skeletogenesis. Little is known about the initial specification of the skeletogenic network as well as the regulation of downstream skeletal effector genes.

1.5.3 The Ophiuroid Skeletogenic Program

Skeletogenesis in brittle star embryos has been recently studied in detail (Dylus et al., 2016). Brittle stars do not form micromeres at the vegetal pole during early development. As in euechinoids, mesenchymal cells (called skeletogenic mesodermal cells) ingress before gastrulation and form two clusters within which spiculogenesis is initiated. Several orthologs of sea urchin skeletogenic genes are expressed in the skeletogenic lineage, including *alx1, tbr, ets1, tgif, erg, hex, delta jun, nk7, p19, p58a* and *p58b*.

A recent dissection of the *Amphiura filiformis* (brittle star) skeletogenic GRN (Dylus et al., 2016) revealed several differences compared to the *S. purpuratus* PMC GRN. The deployment of the skeletogenic program in *A. filiformis* skeletal precursor cells is independent of the double-negative gate. The spatio-temporal expression of the *A. filiformis pmar1* ortholog (*Afi-pplx1*) is almost identical to *Sp-pmar1*, but it does not act as a repressor, and ectopic expression of *Af-pplx* does not re-specify embryonic cells to a skeletogenic fate. Furthermore, the *A. filiformis hesC* ortholog does not repress *Afi-tbr, Afi-ets1/2* and *Afi-delta* expression and is not repressed by *Afi-pplx1*. Several differences were also seen in a set of late regulatory genes, downstream of *alx1, ets1, tbr* and *jun*, that activate the sets of skeletal effector genes. The *A. filiformis* orthologs of *hex, erg* and *tgif* are expressed mostly in similar domains as in *S. purpuratus*, but they are not engaged in an interlocking and persistent, stable loop as in the *S.purpuratus* PMC GRN. During gastrulation, these genes are no longer restricted to the skeletogenic mesenchyme: they are also expressed in the non-skeletogenic mesenchyme. Their activating inputs as well as their order of activation in the skeletogenic mesenchyme is also different. *A. filiformis foxB* and *dri* orthologs do not function in skeletogenesis.

1.5.4 Mesoderm Specification in Asteroids

Sea star larvae do not form any skeletal elements. No micromeres are created during early cleavage, and no skeletogenic mesenchyme is specified. Mesenchymal (non-skeletogenic) cells migrate into the blastocoel only after gastrulation is complete. *Asterina miniata* orthologs of *ets1/2, gatac, otx* and *tbr* are expressed in mesodermal cells (Hinman et al., 2007). A study conducted by McCauley et al. (2010) revealed the expression of other TFs, and their regulatory connections. *Hex, erg, tgif* and *foxN2/3* orthologs are first expressed in the vegetal pole of sea star blastulae and then in distinct endodermal and mesodermal territories by mid-gastrula. The *hesC* ortholog functions as a repressor but does not repress the expression of mesodermal genes, and no *pmar1* ortholog has been identified.

The mesoderm territory is likely established by a recursively wired circuit consisting of *hex, erg* and *tgif*, activated by *tbr*. *Ets1* and *foxN2/3* are downstream of this circuit. The regulatory interactions between *hex, erg* and *tgif* consist primarily of positive feedback loops that serve to ensure stable and robust expression of these genes: this subcircuit is conserved between sea stars and sea urchins despite the lack of skeletogenesis in sea star larvae.

1.5.5 The Evolution of the Larval Endoskeleton in Echinoderms: Insights from Comparative GRN Analysis

Given what we know about the deployment of the skeletogenic network in various echinoderm embryos and the presence or absence of the larval endokeleton, we can begin to answer the bigger question of how this novel structure may have arisen. Adult echinoderms from all five classes have a calcite-based endoskeleton, and the larvae of echinoid, holothuroid and ophiuroid embryos construct an embryonic endoskeleton during development, but asteroids and crinoids do not. The presence of the larval skeleton in echinoids, holothuroids and ophiuroids can be explained by the co-option of the adult skeletogenic program into the embryo in the common ancestor of these groups (Ettensohn, 2013). This view is supported by the fact that several skeletal effector genes that have biomineralization functions in the embryo also function in biomineralizing the spines and test plates of the adult sea urchin (Gao and Davidson, 2008).

A unique feature of modern-day euechinoids, compared to the rest of the echinoderm class, is the precocious specification of the micromeres and the skeletogenic lineage, leading to the construction of a relatively morphologically complex, patterned skeleton. This points to a second co-option event, in which the skeletogenic program that would normally be deployed during or after gastrulation is imported into the micromeres during early cleavage. This second co-option event most likely occurred when the euchinoids split from the cidaroids. This second co-option event is supported by two key observations: first, additional skeletal elements are constructed in the feeding larva by a group of cells not involved in the construction of the early larval endoskeleton, called secondary mesenchyme cells (SMCs), which originate from a group of cells immediately adjacent to the large micromeres that give rise to PMCs. SMCs and PMCs have similar regulatory states, other than the role of Alx1, which is restricted to the PMC lineage (reviewed by Ettensohn (2013)). Second, the skeletogenic network can be artificially activated in the non-skeletogenic mesenchyme (NSM) after gastrulation. If PMCs are removed from the embryo after ingression, the skeletogenic program is activated in NSM cells, and a normally patterned skeleton is formed. A study by Sharma and Ettensohn (2011) showed that the skeletogenic program, when activated in NSM cells, is almost identical to that in PMCs, except for the upstream regulation of the network.

Among echinoderms, the differences in skeletogenic GRN architecture are concentrated

at the initial activation of the skeletogenic program and at the level of the regulation of the skeletogenic effector gene sets. The double-negative gate appears to be a novel feature of euechinoids, while the intermediate regulatory interactions in the network are conserved to various degrees across echinoderms. Regulatory connections to effector genes have not been studied in as much detail in other echinoderms, but based on what is known so far there seems to be little conservation in network architecture at this peripheral level. This is consistent with the "hourglass" model of evolution, in which genes and regulatory connections constituting early and late linkages in a GRN are less conserved than the linkages occurring in the middle.

1.6 Conclusions

While the PMC GRN is among the most comprehensive and detailed networks in any organism, in order to fully understand the genomic control of skeletal morphogenesis, the architecture of this network must be elucidated in greater detail. There are a few missing links in the information we have about the architecture of the network so far. Perturbation analyses that have been used to infer regulatory connections during embryogenesis rely on techniques that knockdown genes very early in development, with no way to conditionally knockdown genes at specific developmental stages or in specific cell types. This limits our understanding of the temporal progression of the network. WMISH screens and transcriptome profiles have provided valuable information about spatio-temporal expression of developmental genes, but cannot capture dynamic changes in regulatory interactions as development proceeds. New technologies such as a CRISPR/Cas9 system utilizing inducible Cas9 variants, and photoactivable morpholinos can be used to dissect temporal changes in the PMC GRN.

We have a fairly comprehensive view of the regulatory and effector genes involved in skeletogenesis as well as putative CRMs mediating the expression of these genes. However, direct regulatory connections between effector genes and upstream TFs have been elucidated for only a handful of genes. Mutational analysis using BAC-GFP constructs and reporter plasmids are time-consuming and laborious, but can be reasonably sped up by injecting several hundred barcoded constructs at once and using qPCR to quantitate expression (Nam and Davidson, 2012). Genome-wide techniques have been successfully used to identify putative CRMs, but connecting distal CRMs to target genes requires additional information that chromosome conformation capture techniques can provide. Additional TF knockdowns followed by RNA-seq can help identify functional linkages between new TFs added to the network and downstream effector genes.

To delineate direct TF-effector gene interactions, binding sites for the various TFs in the network must be identified. Computational methods can be used to find motifs *de novo* from CRM sequences. Genome-wide DNase 1 footprinting can be used to identify TF binding sites within DNase 1 hypersensitive fragments. Protein binding microarrays can

be used to determine consensus binding sites for TFs that can be purified. ChIP-seq can also be used to identify TF binding sites, provided high-quality antibodies against select TFs can be made.

A detailed and comprehensive PMC GRN will help answer fundamental questions around how cell lineages are specified during development and can be used as a model system for comparative studies that shed light on the evolutionary mechanisms that enable the acquisition of novel structures.

Bibliography

Adomako-Ankomah, A. and Ettensohn, C.A., 2011. P58-A and P58-B: novel proteins that mediate skeletogenesis in the sea urchin embryo. *Developmental biology*, **353**(1):81–93. ISSN 1095-564X. doi:10.1016/j.ydbio.2011.02.021.

Adomako-Ankomah, A. and Ettensohn, C.A., 2013. Growth factor-mediated mesodermal cell guidance and skeletogenesis in the sea urchin embryo. *Development*, **140**:4212–4225.

Akasaka, K., Frudakis, T.N., Killian, C.E., George, N.C., Yamasu, K., et al., 1994. Genomic organization of a gene encoding the spicule matrix protein SM30 in the sea urchin Strongylocentrotus purpuratus. *J Biol Chem*, **269**(32):20592–8.

Amore, G. and Davidson, E.H., 2006. cis-Regulatory control of cyclophilin, a member of the ETS-DRI skeletogenic gene battery in the sea urchin embryo. *Dev Biol*, **293**(2):555–64. doi:10.1016/j.ydbio.2006.02.024.

Barsi, J.C., Tu, Q., and Davidson, E.H., 2014. General approach for in vivo recovery of cell type-specific effector gene sets. *Genome Res*, **24**(5):860–8. doi:10.1101/gr.167668.113.

Berger, M.F. and Bulyk, M.L., 2009. Universal protein-binding microarrays for the comprehensive characterization of the DNA-binding specificities of transcription factors. *Nat Protoc*, **4**(3):393–411. doi:10.1038/nprot.2008.195.

Buenrostro, J.D., Wu, B., Chang, H.Y., and Greenleaf, W.J., 2015. ATAC-seq: A Method for Assaying Chromatin Accessibility Genome-Wide. *Curr Protoc Mol Biol*, **109**:21.29.1–9. doi:10.1002/0471142727.mb2129s109.

Cameron, R.A., Mahairas, G., Rast, J.P., Martinez, P., Biondi, T.R., et al., 2000. A sea urchin genome project: sequence scan, virtual map, and additional resources. *Proc Natl Acad Sci U S A*, **97**(17):9514–8. doi:10.1073/pnas.160261897.

Cameron, R.A., Oliveri, P., Wyllie, J., and Davidson, E.H., 2004. cis-Regulatory activity of randomly chosen genomic fragments from the sea urchin. *Gene Expr Patterns*, **4**(2):205–13. doi:10.1016/j.modgep.2003.08.007.

Cary, G.A. and Hinman, V.F., 2017. Echinoderm development and evolution in the post-genomic era. *Dev Biol*. doi:10.1016/j.ydbio.2017.02.003.

Cheers, M.S. and Ettensohn, C.A., 2005. P16 is an essential regulator of skeletogenesis in the sea urchin embryo. *Developmental Biology*, **283**:384–396.

Chuang, C.K., Wikramanayake, A.H., Mao, C.A., Li, X., and Klein, W.H., 1996. Transient appearance of Strongylocentrotus purpuratus Otx in micromere nuclei: cytoplasmic retention of SpOtx possibly mediated through an alpha-actinin interaction. *Dev Genet*, **19**(3):231–7. doi:10.1002/(SICI)1520-6408(1996)19:3⟨231::AID-DVG6⟩3.0.CO;2-A.

Cong, L., Ran, F.A., Cox, D., Lin, S., Barretto, R., et al., 2013. Multiplex genome engi-

neering using CRISPR/Cas systems. *Science*, **339**(6121):819–23. doi:10.1126/science.1231143.

Crawford, G.E., Holt, I.E., Whittle, J., Webb, B.D., Tai, D., et al., 2006. Genome-wide mapping of DNase hypersensitive sites using massively parallel signature sequencing (MPSS). *Genome Res*, **16**(1):123–31. doi:10.1101/gr.4074106.

Damle, S. and Davidson, E.H., 2011. Precise cis-regulatory control of spatial and temporal expression of the alx-1 gene in the skeletogenic lineage of s. purpuratus. *Dev Biol*, **357**(2):505–17. doi:10.1016/j.ydbio.2011.06.016.

Davidson, E.H., 2001. *Genomic regulatory systems: in development and evolution.* Academic Press.

Davidson, E.H., Rast, J.P., Oliveri, P., Ransick, A., Calestani, C., et al., 2002. A genomic regulatory network for development. *Science*, **295**(5560):1669–78. doi:10.1126/science.1069883.

Duloquin, L., Lhomond, G., and Gache, C., 2007. Localized VEGF signaling from ectoderm to mesenchyme cells controls morphogenesis of the sea urchin embryo skeleton. *Development*, **134**(12):2293–302. doi:10.1242/dev.005108.

Dylus, D.V., Czarkwiani, A., Stångberg, J., Ortega-Martinez, O., Dupont, S., et al., 2016. Large-scale gene expression study in the ophiuroid Amphiura filiformis provides insights into evolution of gene regulatory networks. *Evodevo*, **7**:2. doi:10.1186/s13227-015-0039-x.

Erkenbrack, E.M. and Davidson, E.H., 2015. Evolutionary rewiring of gene regulatory network linkages at divergence of the echinoid subclasses. *Proc Natl Acad Sci U S A*, **112**(30):E4075–84. doi:10.1073/pnas.1509845112.

Erwin, D.H., Laflamme, M., Tweedt, S.M., Sperling, E.A., Pisani, D., et al., 2011. The Cambrian conundrum: early divergence and later ecological success in the early history of animals. *Science*, **334**(6059):1091–7. doi:10.1126/science.1206375.

Ettensohn, C.A., 2006. The emergence of pattern in embryogenesis: regulation of beta-catenin localization during early sea urchin development. *Sci STKE*, **2006**(361):pe48. doi:10.1126/stke.3612006pe48.

Ettensohn, C.A., 2009. Lessons from a gene regulatory network: echinoderm skeletogenesis provides insights into evolution, plasticity and morphogenesis. *Development*, **136**(1):11–21. doi:10.1242/dev.023564.

Ettensohn, C.A., 2013. Encoding anatomy: Developmental gene regulatory networks and morphogenesis. *Wiley Periodicals*, **27**:1–27. ISSN 1526-968X. doi:10.1002/dvg.22380.

Ettensohn, C.A. and Dey, D., 2017. KirrelL, a member of the Ig-domain superfamily of

adhesion proteins, is essential for fusion of primary mesenchyme cells in the sea urchin embryo. *Dev Biol*, **421**(2):258–270. doi:10.1016/j.ydbio.2016.11.006.

Ettensohn, C.A., Illies, M.R., Oliveri, P., and De Jong, D.L., 2003. Alx1, a member of the Cart1/Alx3/Alx4 subfamily of Paired-class homeodomain proteins, is an essential component of the gene network controlling skeletogenic fate specification in the sea urchin embryo. *Development*, **130**:2917–2928.

Fire, A., Xu, S., Montgomery, M.K., Kostas, S.A., Driver, S.E., et al., 1998. Potent and specific genetic interference by double-stranded RNA in Caenorhabditis elegans. *Nature*, **391**(6669):806–11. doi:10.1038/35888.

Fuchikami, T., Mitsunaga-Nakatsubo, K., Amemiya, S., Hosomi, T., Watanabe, T., et al., 2002. T- brain homologue (HpTb) is involved in the archenteron induction signals of micromere descendant cells in the sea urchin embryo. *Development*, **129**:5205–5216.

Gao, F. and Davidson, E.H., 2008. Transfer of a large gene regulatory apparatus to a new developmental address in echinoid evolution. *Proc Natl Acad Sci U S A*, **105**(16):6091–6. doi:10.1073/pnas.0801201105.

Hesselberth, J.R., Chen, X., Zhang, Z., Sabo, P.J., Sandstrom, R., et al., 2009. Global mapping of protein-DNA interactions in vivo by digital genomic footprinting. *Nat Methods*, **6**(4):283–9. doi:10.1038/nmeth.1313.

Hinman, V.F., Nguyen, A., and Davidson, E.H., 2007. Caught in the evolutionary act: precise cis-regulatory basis of difference in the organization of gene networks of sea stars and sea urchins. *Dev Biol*, **312**(2):584–95. doi:10.1016/j.ydbio.2007.09.006.

Johnson, D.S., Mortazavi, A., Myers, R.M., and Wold, B., 2007. Genome-wide mapping of in vivo protein-DNA interactions. *Science*, **316**(5830):1497–502. doi:10.1126/science.1141319.

Juliano, C.E., Swartz, S.Z., and Wessel, G.M., 2010. A conserved germline multipotency program. *Development*, **137**(24):4113–26. doi:10.1242/dev.047969.

Kitamura, K., Nishimura, Y., Kubotera, N., Higuchi, Y., and Yamaguchi, M., 2002. Transient activation of the micro1 homeobox gene family in the sea urchin (Hemicentrotus pulcherrimus) micromere. *Dev Genes Evol*, **212**(1):1–10. doi:10.1007/s00427-001-0202-3.

Klein, W.H. and Li, X., 1999. Function and evolution of Otx proteins. *Biochem Biophys Res Commun*, **258**(2):229–33. doi:10.1006/bbrc.1999.0449.

Koga, H., Morino, Y., and Wada, H., 2014. The echinoderm larval skeleton as a possible model system for experimental evolutionary biology. *Genesis*, **52**(3):186–92. doi:10.1002/dvg.22758.

Kurokawa, D., Kitajima, T., Mitsunaga-Nakatsubo, K., Amemiya, S., Shimada, H., et al.,

1999. HpEts, an ets-related transcription factor implicated in primary mesenchyme cell differentiation in the sea urchin embryo. *Mechanisms of development*, **80**:41–52.

Levine, M. and Davidson, E.H., 2005. Gene regulatory networks for development. *Proc Natl Acad Sci U S A*, **102**(14):4936–42. doi:10.1073/pnas.0408031102.

Lin, C.Y. and Su, Y.H., 2016. Genome editing in sea urchin embryos by using a CRISPR/Cas9 system. *Dev Biol*, **409**(2):420–8. doi:10.1016/j.ydbio.2015.11.018.

Livingston, B.T., Killian, C.E., Wilt, F., Cameron, A., Landrum, M.J., et al., 2006. A genome-wide analysis of biomineralization-related proteins in the sea urchin Strongylocentrotus purpuratus. *Developmental Biology*, **300**:335–348.

Logan, C.Y., Miller, J.R., Ferkowicz, M.J., and McClay, D.R., 1999. Nuclear beta-catenin is required to specify vegetal cell fates in the sea urchin embryo. *Development*, **126**(2):345–57.

Makabe, K.W., Kirchhamer, C.V., Britten, R.J., and Davidson, E.H., 1995. Cis-regulatory control of the SM50 gene, an early marker of skeletogenic lineage specification in the sea urchin embryo. *Development*, **121**(7):1957–1970.

Mali, P., Yang, L., Esvelt, K.M., Aach, J., Guell, M., et al., 2013. RNA-guided human genome engineering via Cas9. *Science*, **339**(6121):823–6. doi:10.1126/science.1232033.

Malkov, V.A., Serikawa, K.A., Balantac, N., Watters, J., Geiss, G., et al., 2009. Multiplexed measurements of gene signatures in different analytes using the Nanostring nCounter Assay System. *BMC Res Notes*, **2**:80. doi:10.1186/1756-0500-2-80.

Mann, K., Wilt, F.H., and Poustka, A.J., 2010. Proteomic analysis of sea urchin (Strongylocentrotus purpuratus) spicule matrix. *Proteome Science*, **8**:33.

McCauley, B.S., Weideman, E.P., and Hinman, V.F., 2010. A conserved gene regulatory network subcircuit drives different developmental fates in the vegetal pole of highly divergent echinoderm embryos. *Dev Biol*, **340**(2):200–8. doi:10.1016/j.ydbio.2009.11.020.

McCauley, B.S., Wright, E.P., Exner, C., Kitazawa, C., and Hinman, V.F., 2012. Development of an embryonic skeletogenic mesenchyme lineage in a sea cucumber reveals the trajectory of change for the evolution of novel structures in echinoderms. *Evodevo*, **3**(1):17. doi:10.1186/2041-9139-3-17.

Minokawa, T., Wikramanayake, A.H., and Davidson, E.H., 2005. cis-Regulatory inputs of the wnt8 gene in the sea urchin endomesoderm network. *Dev Biol*, **288**(2):545–58. doi:10.1016/j.ydbio.2005.09.047.

Mortazavi, A., Williams, B.A., McCue, K., Schaeffer, L., and Wold, B., 2008. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat Methods*, **5**(7):621–8. doi:10.1038/nmeth.1226.

Nam, J. and Davidson, E.H., 2012. Barcoded DNA-tag reporters for multiplex cis-regulatory analysis. *PLoS One*, **7**(4):e35934. doi:10.1371/journal.pone.0035934.

Oliveri, P., Carrick, D.M., and Davidson, E.H., 2002. A regulatory gene network that directs micromere specification in the sea urchin embryo. *Dev Biol*, **246**(1):209–28. doi:10.1006/dbio.2002.0627.

Oliveri, P., Davidson, E.H., and McClay, D.R., 2003. Activation of pmar1 controls specification of micromeres in the sea urchin embryo. *Dev Biol*, **258**(1):32–43.

Oliveri, P., Tu, Q., and Davidson, E.H., 2008. Global regulatory logic for specification of an embryonic cell lineage. *Proceedings of the National Academy of Sciences of the United States of America*, **105**(16):5955–62. ISSN 1091-6490. doi:10.1073/pnas.0711220105.

Oulhen, N. and Wessel, G.M., 2016. Albinism as a visual, in vivo guide for CRISPR/Cas9 functionality in the sea urchin embryo. *Mol Reprod Dev*, **83**(12):1046–1047. doi:10.1002/mrd.22757.

Peter, I.S. and Davidson, E.H., 2011. A gene regulatory network controlling the embryonic specification of endoderm. *Nature*, **474**(7353):635–9. doi:10.1038/nature10100.

Peter, I.S. and Davidson, E.H., 2015. *Genomic Control Process: Development and Evolution*. Academic Press.

Rafiq, K., Cheers, M.S., and Ettensohn, C.A., 2012. The genomic regulatory control of skeletal morphogenesis in the sea urchin. *Development*, **139**:579–590.

Rafiq, K., Shashikant, T., McManus, C.J., and Ettensohn, C.A., 2014. Genome-wide analysis of the skeletogenic gene regulatory network of sea urchins. *Development*, **141**(4):950–61. doi:10.1242/dev.105585.

Reich, A., Dunn, C., Akasaka, K., and Wessel, G., 2015. Phylogenomic analyses of Echinodermata support the sister groups of Asterozoa and Echinozoa. *PLoS One*, **10**(3):e0119627. doi:10.1371/journal.pone.0119627.

Revilla-i Domingo, R., Minokawa, T., and Davidson, E.H., 2004. R11: a cis-regulatory node of the sea urchin embryo gene network that controls early expression of SpDelta in micromeres. *Dev Biol*, **274**(2):438–51. doi:10.1016/j.ydbio.2004.07.008.

Revilla-i Domingo, R., Oliveri, P., and Davidson, E.H., 2007. A missing link in the sea urchin embryo gene regulatory network: hesC and the double-negative specification of micromeres. *Proc Natl Acad Sci U S A*, **104**(30):12383–8. doi:10.1073/pnas.0705324104.

Rho, H.K. and McClay, D.R., 2011. The control of foxN2/3 expression in sea urchin embryos and its function in the skeletogenic gene regulatory network. *Development*, **138**(5):937–45. doi:10.1242/dev.058396.

Röttinger, E., Besnardeau, L., and Lepage, T., 2004. A Raf/MEK/ERK signaling pathway is required for development of the sea urchin embryo micromere lineage through

phosphorylation of the transcription factor Ets. *Development*, **131**(5):1075–87. doi: 10.1242/dev.01000.

Röttinger, E., Saudemont, A., Duboc, V., Besnardeau, L., McClay, D., et al., 2008. FGF signals guide migration of mesenchymal cells, control skeletal morphogenesis [corrected] and regulate gastrulation during sea urchin development. *Development*, **135**(2):353–65. doi:10.1242/dev.014282.

Sea Urchin Genome Sequencing Consortium, Sodergren, E., Weinstock, G.M., Davidson, E.H., Cameron, R.A., et al., 2006. The genome of the sea urchin Strongylocentrotus purpuratus. *Science*, **314**(5801):941–52. doi:10.1126/science.1133609.

Sharma, T. and Ettensohn, C.A., 2010. Activation of the skeletogenic gene regulatory network in the early sea urchin embryo. *Development*, **137**(7):1149–57. doi:10.1242/dev.048652.

Sharma, T. and Ettensohn, C.A., 2011. Regulative deployment of the skeletogenic gene regulatory network during sea urchin development. *Development*, **138**(12):2581–90. doi: 10.1242/dev.065193.

Smith, J., 2008. A protocol describing the principles of cis-regulatory analysis in the sea urchin. *Nat Protoc*, **3**(4):710–8. doi:10.1038/nprot.2008.39.

Smith, J., Theodoris, C., and Davidson, E.H., 2007. A gene regulatory network subcircuit drives a dynamic pattern of gene expression. *Science*, **318**(5851):794–7. doi:10.1126/science.1146524.

Summerton, J. and Weller, D., 1997. Morpholino antisense oligomers: design, preparation, and properties. *Antisense Nucleic Acid Drug Dev*, **7**(3):187–95. doi:10.1089/oli.1.1997.7.187.

Sun, Z. and Ettensohn, C.A., 2017. TGF-B sensu stricto signaling regulates skeletal morphogenesis in the sea urchin embryo. *Dev Biol*, **421**(2):149–160. doi:10.1016/j.ydbio.2016.12.007.

Sweet, H.C., Gehring, M., and Ettensohn, C.A., 2002. LvDelta is a mesoderm-inducing signal in the sea urchin embryo and can endow blastomeres with organizer-like properties. *Development*, **129**(8):1945–55.

Telford, M.J., Lowe, C.J., Cameron, C.B., Ortega-Martinez, O., Aronowicz, J., et al., 2014. Phylogenomic analysis of echinoderm class relationships supports Asterozoa. *Proc Biol Sci*, **281**(1786). doi:10.1098/rspb.2014.0479.

Tu, Q., Cameron, R.A., and Davidson, E.H., 2014. Quantitative developmental transcriptomes of the sea urchin Strongylocentrotus purpuratus. *Dev Biol*, **385**(2):160–7. doi: 10.1016/j.ydbio.2013.11.019.

Tu, Q., Cameron, R.A., Worley, K.C., Gibbs, R.a., and Davidson, E.H., 2012. Gene struc-

ture in the sea urchin Strongylocentrotus purpuratus based on transcriptome analysis. *Genome research*, **22**(10):2079–87. ISSN 1549-5469. doi:10.1101/gr.139170.112.

Wahl, M.E., Hahn, J., Gora, K., Davidson, E.H., and Oliveri, P., 2009. The cis-regulatory system of the tbrain gene: Alternative use of multiple modules to promote skeletogenic expression in the sea urchin embryo. *Dev Biol*, **335**(2):428–41. doi:10.1016/j.ydbio.2009.08.005.

Weitzel, H.E., Illies, M.R., Byrum, C.A., Xu, R., Wikramanayake, A.H., et al., 2004. Differential stability of beta-catenin along the animal-vegetal axis of the sea urchin embryo mediated by dishevelled. *Development*, **131**(12):2947–56. doi:10.1242/dev.01152.

Wikramanayake, A.H., Huang, L., and Klein, W.H., 1998. beta-Catenin is essential for patterning the maternally specified animal-vegetal axis in the sea urchin embryo. *Proc Natl Acad Sci U S A*, **95**(16):9343–8.

Wilt, F.H. and Ettensohn, C.A., 2007. The morphogenesis and biomineralization of the sea urchin larval skeleton. *Handbook of Biomineralization: Biological Aspects and Structure Formation (ed. E. Bauerlein)*, pages 183–210.

Wu, S.Y. and McClay, D.R., 2007. The Snail repressor is required for PMC ingression in the sea urchin embryo. *Development (Cambridge, England)*, **134**(6):1061–70. ISSN 0950-1991. doi:10.1242/dev.02805.

Wu, S.Y., Yang, Y.P., and McClay, D.R., 2008. Twist is an essential regulator of the skeletogenic gene regulatory network in the sea urchin embryo. *Dev Biol*, **319**(2):406–15. doi:10.1016/j.ydbio.2008.04.003.

Yajima, M. and Wessel, G.M., 2011. Small micromeres contribute to the germline in the sea urchin. *Development*, **138**(2):237–43. doi:10.1242/dev.054940.

Yajima, M. and Wessel, G.M., 2012. Autonomy in specification of primordial germ cells and their passive translocation in the sea urchin. *Development*, **139**(20):3786–94. doi:10.1242/dev.082230.

Yamasu, K. and Wilt, F.H., 1999. Functional organization of DNA elements regulating SM30alpha, a spicule matrix gene of sea urchin embryos. *Dev Growth Differ*, **41**(1):81–91.

Yamazaki, A., Kidachi, Y., Yamaguchi, M., and Minokawa, T., 2014. Larval mesenchyme cell specification in the primitive echinoid occurs independently of the double-negative gate. *Development*, **141**(13):2669–79. doi:10.1242/dev.104331.

Chapter 2

Genome-wide Identification of Skeletal Morphogenesis Genes

This chapter consists of the paper authored by Kiran Rafiq (co-first author), Tanvi Shashikant (co-first author), C. Joel McManus and Charles A Ettensohn, published in Development in 2014. My contribution to this paper was: isolation of PMCs and other cells, extraction and QC of RNA from PMCs and other cells, RNA-seq analysis of sequence data obtained from PMCs and other cells as well as analysis of RNA-seq data obtained from Ets1, Alx1 and U0126 knockdown samples.

2.1 Abstract

A central challenge of developmental biology is to understand the transformation of genetic information into morphology. Elucidating the connections between genes and anatomy will require model morphogenetic processes that are amenable to detailed analysis of cell/tissue behaviors and to systems-level approaches to gene regulation. The formation of the calcified endoskeleton of the sea urchin embryo is a valuable experimental system for developing such an integrated view of the genomic regulatory control of morphogenesis. A transcriptional gene regulatory network (GRN) that underlies the specification of skeletogenic cells (primary mesenchyme cells, or PMCs) has recently been elucidated. In this study, we carried out a genome-wide analysis of mRNAs encoded by effector genes in the network. We used RNA-seq to identify 420 transcripts differentially expressed by PMCs at the onset of gastrulation, when these cells undergo a striking sequence of behaviors that drives skeletal morphogenesis. Our analysis expanded by almost an order of magnitude the number of known (and candidate) downstream effectors that directly mediate skeletal morphogenesis. We carried out genome-wide analysis of (1) functional targets of Ets1 and Alx1, two pivotal, early transcription factors in the

PMC GRN, and (2) functional targets of MAPK signaling, a pathway that plays an essential role in PMC specification. These studies identified transcriptional inputs into >200 PMC effector genes. Our work establishes a framework for understanding the genomic regulatory control of a major morphogenetic process and has important implications for reconstructing the evolution of biomineralization in metazoans.

2.2 Introduction

The progressive changes in form that characterize embryogenesis are encoded in the genome. The properties of cells that drive these changes in form, like other specialized cellular properties, arise as a consequence of differential gene expression. Programs of differential gene expression can be viewed as dynamic networks of regulatory genes (genes that encode transcription factors, or TFs), and the cis- regulatory DNA elements to which TFs bind. Such gene regulatory networks (GRNs) are proving to be powerful tools for analyzing cell specification and the evolution of development (Stathopoulos and Levine, 2005; Davidson, 2010; Peter and Davidson, 2011; Van Nostrand and Kim, 2011; Wunderlich and DePace, 2011). A current limitation of this conceptual approach to development, however, is that we have a poor understanding of the connections between transcriptional networks and the morphogenetic processes that build tissues and organs. A marriage of regulatory network biology and morphogenesis will require experimental models that are amenable both to systems-level approaches and to detailed analysis of morphogenetic mechanisms. Integrating transcriptional networks and morphogenesis will also be crucial in an evolutionary context, i.e. for understanding how evolutionary modifications to genetic programs have supported changes in animal anatomy (Ettensohn, 2013).

The endoskeleton of the sea urchin embryo provides an opportunity to elucidate the genetic circuitry that underlies the formation of a major anatomical feature. The skeleton is a biomineral composed of calcium carbonate (in the form of calcite) and small amounts of occluded proteins. It is secreted by primary mesenchyme cells (PMCs), a population of cells derived from the large micromeres (LMs) of the cleavage stage embryo. During gastrulation, PMCs undergo a sequence of morphogenetic behaviors that includes ingression (epithelial-mesenchymal transition), directional migration, cell-cell fusion and biomineral formation (Wilt and Ettensohn, 2007; Ettensohn, 2013). These cellular behaviors have been analyzed in detail in living embryos (Gustafson and Wolpert, 1967; Malinda et al., 1995; Miller and McClay, 1995; Guss and Ettensohn, 1997; Peterson and McClay, 2003; Hodor and Ettensohn, 2008; Adomako-Ankomah and Ettensohn, 2013). The skeleton has several important functions; it influences the shape, orientation, swimming and feeding of the larva (Pennington and Strathmann, 1990; Hart and Strathmann, 1994), and its growth during larval development is responsive to environmental cues (Adams et al., 2011).

A GRN that underlies skeletogenic specification is activated in the LM-PMC lineage by localized maternal factors (Emily-Fenouil et al., 1998; Wikramanayake et al., 1998; Logan

et al., 1999; Weitzel et al., 2004; Ettensohn, 2006). These maternal inputs function cell-autonomously to drive the zygotic expression of a small number of lineage-specific TFs, including Ets1 (Kurokawa et al., 1999) and Alx1 (Ettensohn et al., 2003). Early TFs in the GRN engage additional layers of regulatory genes, and various feedback and feed-forward interactions stabilize the network and drive it forward (Oliveri et al., 2008). Although considerable information is available concerning interactions among regulatory genes in the network, we have a very limited understanding of the downstream effector genes that execute skeletal morphogenesis and their transcriptional control.

In previous studies, we used an *in situ* hybridization screen to identify candidate effector genes in the PMC GRN and analyzed the developmental functions and regulatory control of several of these genes (Illies et al., 2002; Cheers and Ettensohn, 2005; Livingston et al., 2006; Adomako-Ankomah and Ettensohn, 2011; Rafiq et al., 2012; Adomako-Ankomah and Ettensohn, 2013). Here, we expand this analysis to a genome-wide level by carrying out an RNA-seq based analysis of effector genes in the PMC GRN. We increase by approximately an order of magnitude the number of known PMC-enriched transcripts. The great majority of these encode effector proteins, many with known or predicted functions, whereas others encode newly identified, PMC-specific TFs. We find that Ets1 and/or Alx1 provide essential regulatory inputs into >50% of the genes differentially expressed by PMCs at the early gastrula stage, pointing to the pivotal role of these TFs in controlling the cell- specific identity of PMCs. Genome-wide mRNA profiling of embryos treated with the MEK inhibitor U0126, which blocks PMC specification by inhibiting the phosphorylation of Ets1 (Fernandez-Serra et al., 2004; Röttinger et al., 2004), reveals that the PMC GRN is a major target of MAPK signaling during early embryogenesis and shows that Ets1 and Alx1 are key mediators of MAPK inputs into the GRN. Overall, this work greatly expands our understanding of the architecture and regulation of the PMC GRN and extends the utility of this experimental system as a model for developing an integrated view of the genomic regulatory control of morphogenesis.

2.3 Results

2.3.1 RNA-seq analysis of mRNAs differentially expressed by PMCs at the onset of gastrulation

Our previous work focused on a subset of highly abundant, PMC-enriched transcripts (Rafiq et al., 2012). To obtain a more global perspective, we used RNA-seq to compare the abundance of transcripts in PMCs and a non-PMC fraction at the early mesenchyme blastula stage 24 hours post-fertilization (hpf). At this stage of development, PMCs are the only cells that have ingressed into the blastocoel. Thus, we enriched PMCs by isolating basal lamina bags from embryos at this stage (Harkey and Whiteley, 1980). Most effector genes in the PMC GRN are expressed at 24 hpf (Rafiq et al., 2012).

We compared the expression of 21,000 distinct *S. purpuratus* transcripts (Tu et al., 2012) in PMCs and the non-PMC fraction. Most mRNAs were expressed at similar levels (Fig. 2.1; $R^2 = 0.91, p < 2 \times 10^{16}$. Cuffdiff analysis identified 420 transcripts with expression levels that differed significantly in the PMC and non- PMC samples (supplementary material Tables S1 and S2). All but five of these mRNAs were more abundant in PMCs than in the non- PMC sample. We refer to the genes that encode this collection of 420 mRNAs as the differentially expressed (DE) gene set. A summary of information concerning the 420 DE genes is presented in supplementary material Table S1 and quantitative expression values for all *S. purpuratus* transcripts from the RNA-seq analysis are provided in supplementary material Table S2. Overall, RNA- seq-based gene expression profiling increased by approximately an order of magnitude the number of known PMC-enriched mRNAs and therefore provided a far more complete picture of the output of this transcriptional network than was previously available.

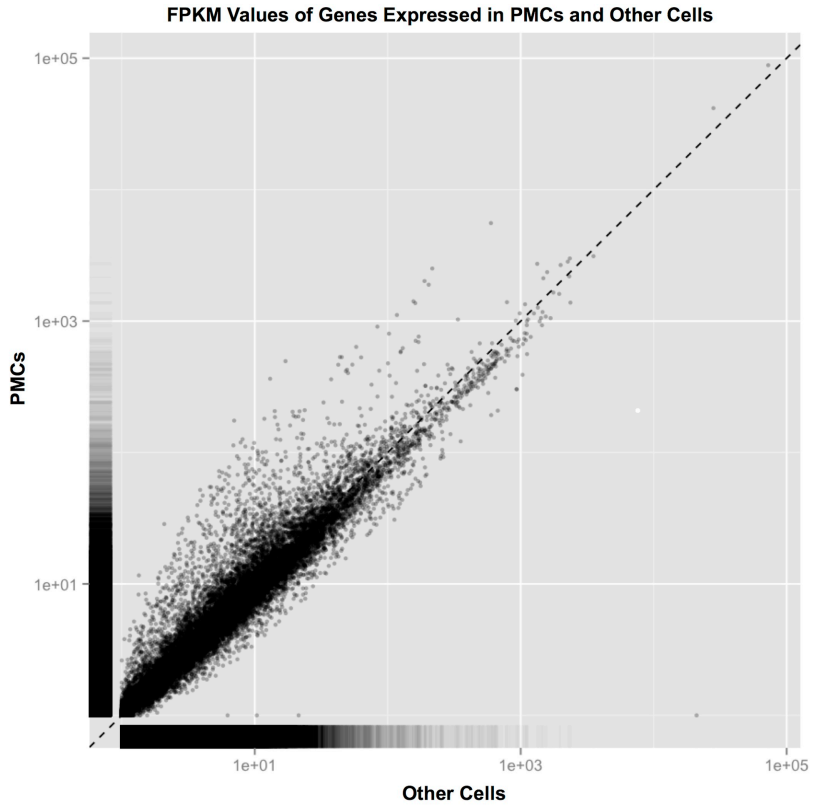


Figure 2.1: Linear scatter plot of FPKM values derived from RNA-seq analysis of genes expressed at 24 hpf in PMCs and a non-PMC fraction isolated from *S. purpuratus* embryos. FPKM values shown are the means of two biological replicates and range from <1 to 5564 in purified PMCs and <1 to 3524 in the non-PMC fraction ('other cells'), for the 21,000 transcripts detected. $R^2 = 0.91, p < 2 \times 10^{16}$

To assess the completeness of our analysis (i.e. the false-negative rate), we examined a set of 36 mRNAs that were previously reported to be restricted to PMCs at this developmental stage, based on a WMISH screen and literature survey [see table S2 in the supplementary material of Rafiq et al. (2012)]. Of these transcripts, only 4/36 (11%) were not found in the collection of DE genes. The four mRNAs that were not identified (*ctd, p19, sm37 and stomatin*) all yielded FPKM values in PMCs that were higher than in the non-PMC cell fraction, but the data failed to meet the significance criteria of the Cuffdiff analysis. This sampling indicates that, although the collection of DE genes is not exhaustive, it is likely to have captured the great majority of transcripts that are differentially expressed by PMCs at the mesenchyme blastula stage.

Because the significance threshold of the Cuffdiff analysis was relatively stringent (estimated false discovery rate=0.05), it seems likely that the DE set contains few false positives. We took two approaches to further assess the frequency of false positives in the DE gene set. First, we examined 50 genes chosen at random from those *S. purpuratus* genes annotated with Gene Ontology (GO) terms associated with metabolism, DNA replication, protein translation and other likely housekeeping functions. None of these genes was found to be differentially expressed in our analysis. Second, we used WMISH to analyze the expression patterns of 41 DE genes (these were selected because examination of the predicted gene products suggested a possible role in skeletal morphogenesis, as discussed below). WMISH analysis confirmed that 25 of these mRNAs were enriched in PMCs at the mesenchyme blastula stage (and, in most cases, at later developmental stages) (Fig. 2.2).

The remaining WMISH probes, most of which were directed against low-abundance transcripts, showed uniformly low levels of staining. Based on analysis of $>$200 PMC-enriched mRNAs (Rafiq et al., 2012), we found that the threshold for WMISH detection was an FPKM value of 5-10 in whole embryo samples, which corresponded to 4-7 transcripts/PMC (assuming 32 PMCs/embryo). In general, RNA- seq data agreed well with WMISH analysis, i.e. transcripts that were (1) relatively abundant as indicated by a high FPKM value and (2) highly enriched in PMCs as reflected by a high log2-fold difference (supplementary material Table S1), yielded robust WMISH patterns that were restricted to PMCs at the mesenchyme blastula stage.

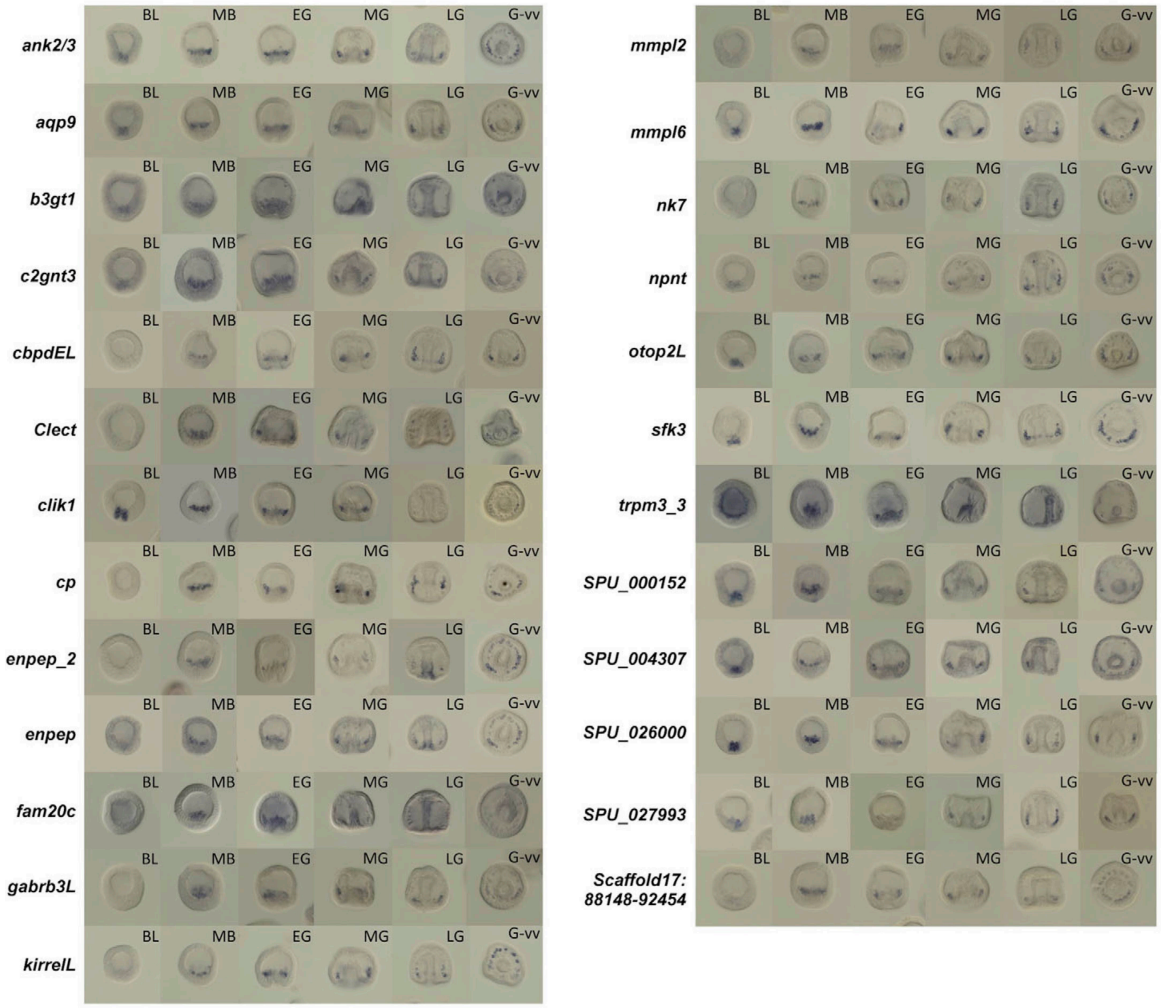


Figure 2.2: WMISH analysis of mRNAs differentially expressed in PMCs. Lateral views are shown for all stages and gastrula stage embryos are also viewed along the animal-vegetal axis (far right panels; ventral side down). The gene names shown are those assigned in SpBase. For genes that have not been assigned descriptive names, SPU identifiers (also from SpBase) are shown. Also included is a previously unannotated gene, located on Scaffold17:88148- 92454 (v.3.1), which encodes a member of a family of PMC-specific Ig/TM proteins. BL, blastula; MB, mesenchyme blastula; EG, early gastrula; MG, mid- gastrula; LG, late gastrula; G-vv, mid- to late gastrula-vegetal view.

2.3.2 Characterization of DE genes

Most of the 420 DE genes were expressed at relatively low levels at the mesenchyme blastula stage (<10 transcripts/PMC, assuming 16 PMCs/embryo at this stage; Supp Fig. 2.1). Only 39 (∼10%) of the PMC-enriched mRNAs were expressed at levels greater than 25

transcripts/PMC. Not surprisingly, many of these abundant transcripts encoded biomineralization proteins, including three members of the Msp130 family (Msp130, Msp130r1 and Msp130r2), six spicule matrix proteins [Sm20 (Clect_14), Sm29, Sm30, Sm49 (C-lectin/PMC1), Sm32/50 and C-lectin], and other biomineralization proteins such as P16, P16rel2 (Hypp_2998), P58A (FcgbpL) and carbonic anhydrase (Cara7LA) (Livingston et al., 2006; Rafiq et al., 2012).

The temporal expression profiles of the 420 DE genes were extracted from transcriptome profiling data of Tu et al. (2012) and analyzed by hierarchical clustering (Fig. 2.3). This analysis revealed subsets of DE genes with coordinated temporal expression profiles, including (from top to bottom in Fig. 2.3): (1) genes with high levels of maternal expression; (2) a small set of genes that showed a sharp spike in expression during late cleavage (10 hpf); (3) genes that showed maximal expression during late gastrulation and post-gastrula stages (48-72 hours); and (4) a major class of genes (almost half) that were expressed at very low levels early in development, peaked in expression during the late blastula-gastrula stages (18-40 hpf), and then declined in expression (genes in the lower half of Fig. 2.3).

To gain insight into the possible roles of DE genes, we first examined the functional assignments of these genes as annotated in SpBase (Cameron et al., 2009). Forty-six percent of the DE genes have been assigned to functional categories based on hand annotations (Sodergren et al., 2006) or primary GO terms derived by blast2go (Tu et al., 2012). Fig. 2.4 shows the distribution of these functional classes. Consistent with the skeletogenic function of PMCs, one of the largest classes of DE genes was the biomineralization set. It seems likely that many of the DE genes without functional annotations (as well as genes currently annotated as 'novel') also encode biomineralization-related proteins. In addition, some genes in the 'calcium toolkit', 'kinase' and 'metalloprotease' classes play important roles in biomineral formation. As an independent means of assessing the subset of DE genes with functions in biomineralization, we examined a set of $\sim$ 200 proteins identified in a recent proteomic analysis of partially purified embryonic spicules (Mann et al., 2010) and identified 62 gene products that were common to the two sets (supplementary material Table S1). Adhesion-related proteins constituted another sizable functional category, and many of these proteins are likely to be involved in PMC migration and/or fusion (see below).

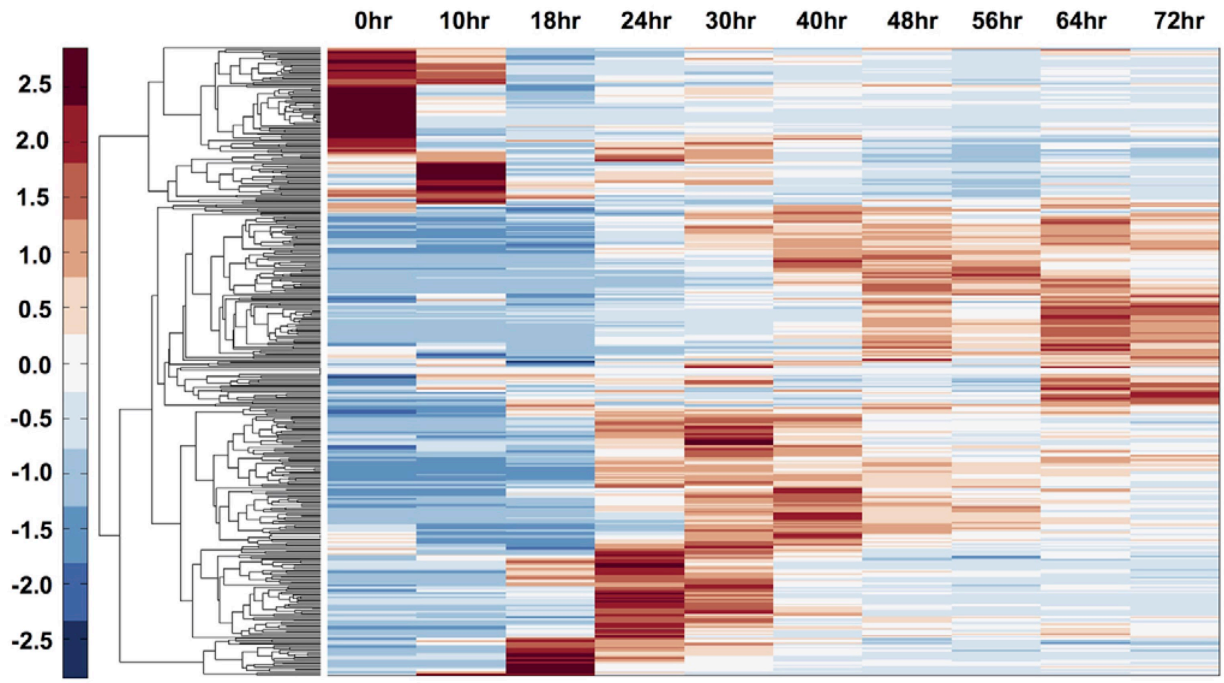


Figure 2.3: Hierarchical clustering of temporal expression patterns of genes differentially expressed in PMCs. The temporal expression profiles of 420 genes differentially expressed in PMCs were obtained from the RNA-seq data of Tu et al. (2012), available on SpBase. Each gene is represented by a single row and each time point by a single column. The color scale ranges from deep red (2.5-fold higher than mean expression) to deep blue (2.5-fold lower than mean expression). White indicates the mean expression value. For reference, 24, 48 and 72 hpf correspond to the mesenchyme blastula, late gastrula and late prism stages, respectively.

We examined the complete set of 368 annotated TFs in the *S. purpuratus* genome and found that more than half were detectable in basal lamina bag-purified PMCs at levels >1 transcript/cell. Several TFs in this set, however, including *gcm, foxa, scl* and *gataC*, are restricted to cell types other than PMCs at the mesenchyme blastula stage (Ransick and Davidson, 2006; Croce and McClay, 2010; Flynn et al., 2011) and their identification in our analysis reflected the low level of contamination of the bag preparations with cell types other than PMCs. When we restricted our analysis to the set of 420 DE genes, we identified only 11 TFs. One of these (*evi1*) was expressed at lower levels in PMCs than in other cells; the other ten TF mRNAs (*alx1, alx4, cebpa, foxB, foxO, mef2, mitf, nk7, smad2/3* and *tbr*) were enriched in PMCs relative to the non-PMC population to varying degrees, ranging from 3.5-fold (*foxO*) to 15-fold (*nk7*). WMISH data are available for seven of these genes [*alx1* (Ettensohn et al., 2003), *alx4* (Rafiq et al., 2012), *foxb* (Minokawa et al., 2004), *foxO* (Tu et al., 2006), *smad2/3* (Poustka et al., 2007) *tbr* (Fuchikami et al., 2002) and *nk7* (this study, Fig. 2.2] and in all cases confirms that expression is enriched in PMCs at the

mesenchyme blastula stage.

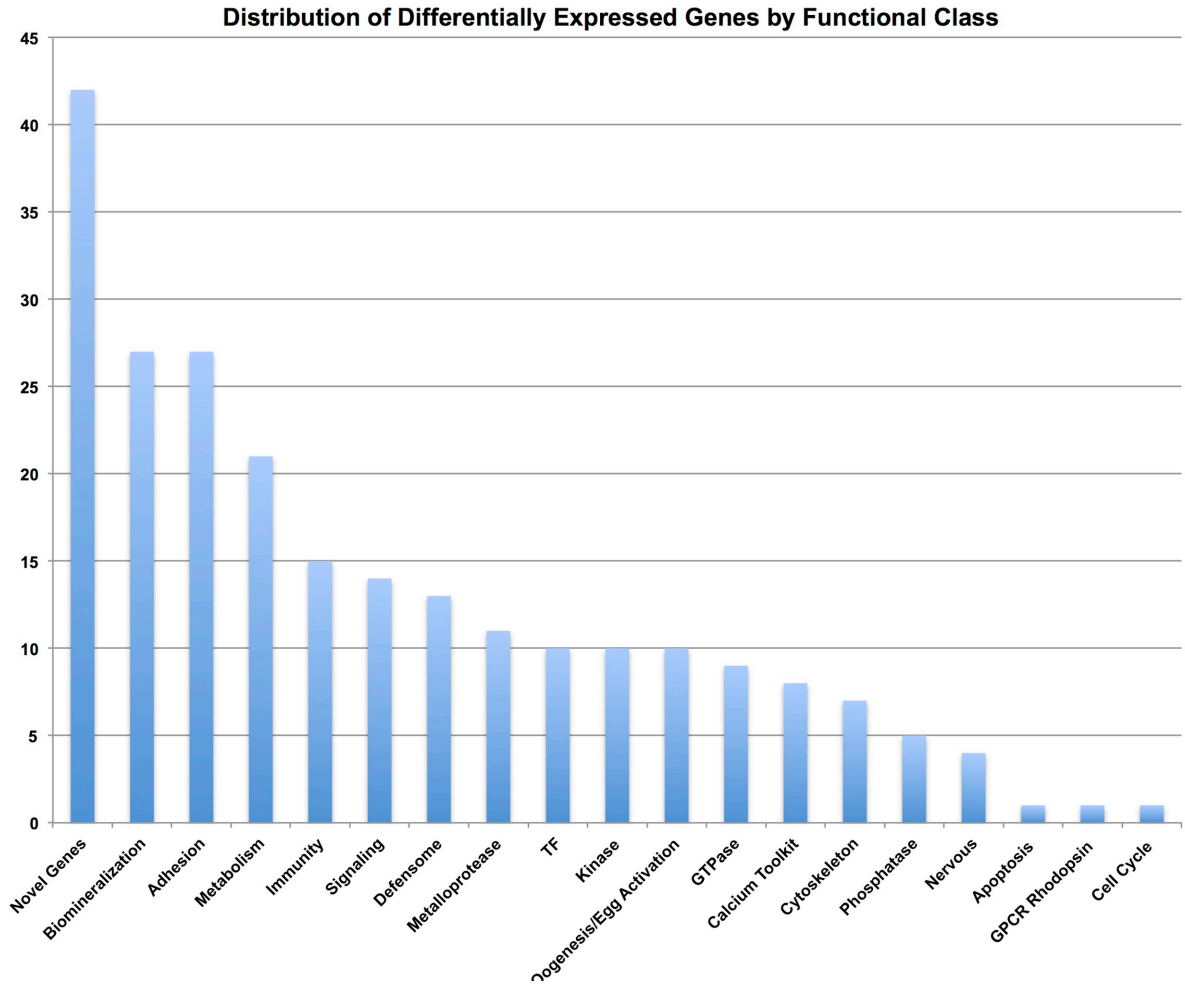


Figure 2.4: Distribution of DE genes by functional class. The distribution is based on the primary functional assignments of DE genes in their public annotations (SpBase). Functional assignments are based on hand annotation (Sodergren et al., 2006) and, where lacking, on primary GO terms derived by blast2go (Tu et al., 2012). Out of the 420 DE genes, 194 have been assigned to functional categories. Novel genes, biomineralization genes and adhesion genes constitute almost half of this set. The y-axis indicates the number of genes in each functional class.

Hand curation of the set of DE genes revealed many new candidate effectors of skeletal morphogenesis, some of which are highlighted below.

Biomineralization proteins

Transport/channel proteins

The deposition of $CaCO_3$ by PMCs is associated with the uptake of Ca^{2+} and HCO_3^- ions from the blastocoel (Stumpp et al., 2012). We identified five solute carrier proteins (members of the Slc4, Slc5, Slc10, Slc24 and Slc26 families) that might mediate these transport functions.

Secreted metalloproteases

Metalloprotease inhibitors reversibly block spiculogenesis by PMCs *in vivo* and *in vitro* (Roe et al., 1989; Ingersoll and Wilt, 1998). The DE genes include a suite of four matrix metalloprotease genes, arranged in tandem on a single chromosome, that are likely to encode the relevant enzymes. Fig. 2.2 shows WMISH data for two of these genes (*mmpl2* and *mmpl6*).

Fam20C

This secreted kinase was recently shown to phosphorylate extracellular biomineralization proteins in vertebrates (Ishikawa et al., 2012; Tagliabracci et al., 2012).

Otopetrin

Otopetrin (*otop2L*) is a transmembrane (TM) protein essential for the development of otoliths/otoconia, which are extracellular calcium carbonate-containing crystals that mediate vestibular mechanosensory function in vertebrates (Hurle et al., 2003; Hughes et al., 2004; Söllner et al., 2004).

Adhesion/migration proteins

Nephronectin

During their migration, PMCs interact with extracellular matrix (ECM) fibers that contains the sea urchin ortholog of vertebrate Frem2 (Hodor et al., 2000). Frem2 and related proteins have been implicated in epithelium-mesenchyme adhesion and mutations in these genes underlie Fraser's syndrome (Smyth and Scambler, 2005). Frem proteins are required for the proper incorporation of nephronectin, an integrin ligand required for kidney morphogenesis, into the ECM (Kiyozumi et al., 2012). The sea urchin ortholog of nephronectin (*npnt*) is expressed selectively by PMCs during gastrulation and might play an important role in PMC-substrate interactions.

Adhesion receptors

PMCs selectively express several type I TM proteins with variable numbers of extracellular Ig, Egf, Lrr and Fn3 repeats, an organization which suggests that these proteins

might function as adhesion receptors. Examples include Lrr/Igr_10, Fn3/Egff_1 and Fn3f_9.

Aquaporin-9

One abundant, PMC-specific transcript (*aqp9*) encodes a member of the aquaporin family of TM, water channel proteins, which have recently been implicated in regulating the protrusive activity and migration of cancer cells (Verkman, 2011).

Cell-cell fusion proteins

The dynamics of PMC fusion have been analyzed *in vivo* (Hodor and Ettensohn, 1998), but molecules that mediate fusion have not been identified. In *Drosophila*, TM proteins with multiple extracellular Ig domains (Sns, Rst and Duf) are required for myoblast interactions prior to fusion (Abmayr and Pavlath, 2012). We have identified four PMC-specific, type I TM proteins with multiple extracellular Ig repeats that are the closest relatives of Sns/Rst/Duf in the sea urchin genome and strong candidates for regulators of PMC fusion. WMISH data for three of these genes (*kirre1L*, *SPU_026000* and *Scaffold17:88148-92454*) are shown in Fig. 2.2.

2.3.3 Transcriptional inputs into DE genes

To identify regulatory inputs into the 420 DE genes, we used RNA-seq to analyze changes in gene expression following knockdown of Ets1 or Alx1 (supplementary material Tables S3 and S4). Ets1 and Alx1 are pivotal early TFs in the PMC GRN (Kurokawa et al., 1999; Ettensohn et al., 2003). RNA-seq was used to profile gene expression in controls and morphants at 28-30 hpf (early gastrula). We chose this stage because the severity of the morphant phenotypes could be unambiguously scored (see Materials and Methods) and because an earlier, more limited, analysis of gene expression changes in Ets1 and Alx1 morphants was carried out at this stage (Rafiq et al., 2012). Most effector genes in the PMC GRN are robustly expressed at 28-30 hpf (Rafiq et al., 2012) (Fig. 2.3).

To assess the reliability of our RNA-seq analysis, we compared QPCR data from a previous study that examined the effects of Ets1 and Alx1 knockdowns on the expression of $\sim$20 effector genes (Rafiq et al., 2012) with Nanostring and RNA-seq-based expression data obtained in the present study for the same set of genes at the same developmental stage. This analysis showed that effects of knockdowns on gene expression were highly reproducible across these experiments, which were carried out using embryos derived from three different male-female matings and which used three different methods of transcript quantification (Supp. Fig. 2.2).

RNA-seq-based gene expression profiling showed that 223/420 DE genes (53%) were significantly affected by knockdown of Ets1 and/or Alx1 (Fig. 2.5; supplementary material Table S1). This demonstrated the pivotal role of these TFs in controlling the cell- specific identity of PMCs. Of the DE genes with inputs from Ets1 or Alx1, most (144/223, or ∼65%) were downregulated in both classes of morphants.

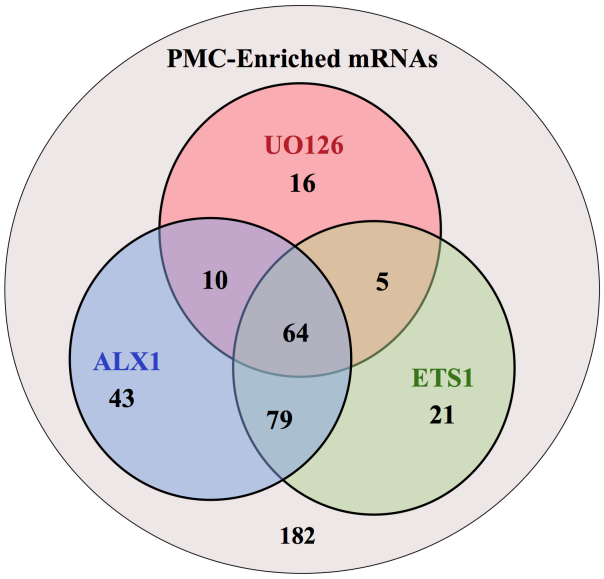


Figure 2.5: Venn diagram showing overlapping distributions of genes affected by Ets1 knockdown, Alx1 knockdown, or U0126 treatment among the 420 genes differentially expressed by PMCs. More than half (223/420, 53%) of DE genes are affected by knockdown of Ets1 and/or Alx1; the great majority of these inputs (∼90%) are positive. Of these 223 DE genes, ∼65% (144/223) are affected in both classes of morphants. 101 DE genes are sensitive to U0126, a number that includes more than half of all U0126 targets genome-wide. Most of the U0126-sensitive DE genes have inputs (direct or indirect) from Ets1 and/or Alx1.

We compared the temporal expression profiles of two cohorts of DE genes: (1) those that were affected both by Ets1 and Alx1 knockdowns, and (2) those that were not regulated by either TF (Fig. 2.6). Hierarchical clustering revealed that the Ets1/Alx1- regulated gene set contained few genes that exhibited high levels of maternal transcripts and, more strikingly, the majority of these genes had a strong spike in expression between the late blastula and mid- gastrula stages (18-30 hpf) (Fig. 2.6 A). By contrast, DE genes that were not regulated by Ets1 or Alx1 showed a much broader distribution of expression patterns, with peak expression levels distributed relatively evenly across all developmental stages, and many genes showed high levels of maternal expression (Fig. 2.6B).

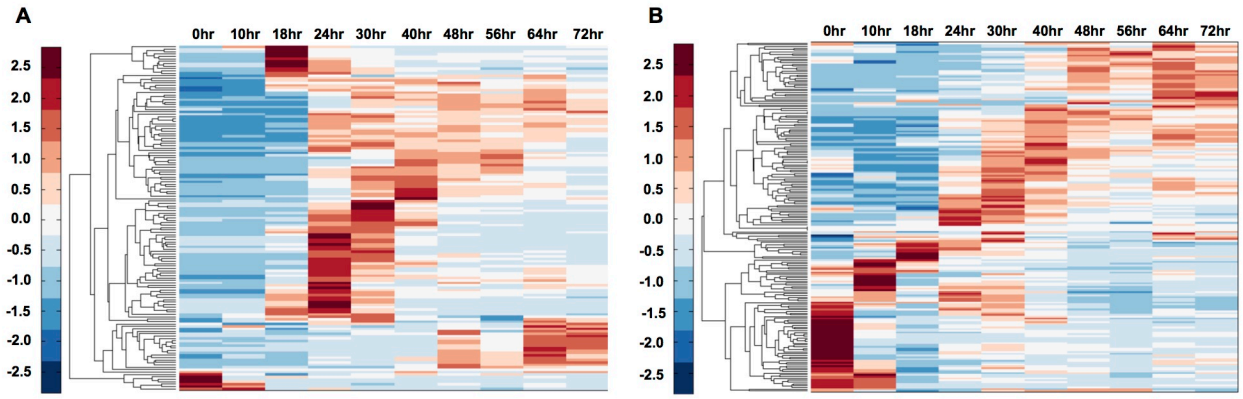


Figure 2.6: Distinct temporal gene expression profiles of Ets1/Alx1 co-regulated targets and non-target genes in the DE set. Hierarchical clustering of the temporal expression profiles of (A) 143 DE genes that are sensitive both to Ets1 and Alx1 knockdown and (B) 198 DE genes that are not regulated by either Ets1 or Alx1 (see Fig.2.3 legend for details). The Ets1/Alx1-regulated gene set contains few genes that exhibit high levels of maternal transcripts and most genes show maximal expression between the late blastula and mid-gastrula stages (18-30 hpf). By contrast, DE genes that are not affected by Ets1 or Alx1 knockdowns show a much broader distribution of temporal expression patterns, including many cases of high maternal expression.

In parallel studies, we carried out RNA-seq transcriptional profiling of 28- to 30-hour embryos that had been treated from the 2-cell stage with U0126, a MEK inhibitor that blocks PMC specification by inhibiting the phosphorylation of Ets1 (Fernandez-Serra et al., 2004; Röttinger et al., 2004). Genome-wide, we identified 180 transcripts that showed significant changes of expression in U0126-treated embryos (supplementary material Table S5). Remarkably, the majority of these mRNAs (101/180, 56%) were also DE genes, suggesting that the PMC GRN is the principal target of MAPK signaling during early embryogenesis (Fig. 2.5). All 101 of the DE mRNAs significantly affected by U0126 treatment were downregulated in the presence of the inhibitor. Of these, the large majority (83%) were also regulated by Ets1 and/or Alx1 (67% were affected in both classes of morphants), pointing to these two TFs as key mediators of MAPK inputs into the GRN (Fig.2.5).

2.3.4 Non-DE genes

The expression of many (>1500) genes not in the DE set was also significantly affected by Ets1/Alx1 knockdowns and/or U0126 treatment (see supplementary material Tables S3-S5). These might include early targets of Ets1/MAPK in the non-skeletogenic mesoderm (NSM) (Fernandez-Serra et al., 2004; Röttinger et al., 2004), but most of the gene expression changes are probably indirect and reflect the additive effects of complex tissue interactions. As an initial step in analyzing these targets, we focused primarily on the

suite of all regulatory genes, as these are well annotated and WMISH data are available for almost all regulatory genes expressed at detectable levels during embryogenesis.

Oliveri et al. (Oliveri et al., 2008) showed that one function of *alx1* in the LM progeny is to repress *gcm*, a regulatory gene ordinarily expressed by adjacent, presumptive pigment cells. Our RNA-seq analysis confirmed an increase in *gcm* expression in Alx1 morphants. We found that Alx1 knockdown resulted in a significant upregulation of 23 other regulatory genes. WMISH data are available for eight of these genes, and a surprisingly large fraction (6/8) are expressed selectively in the NSM during normal development. Four genes (*scl, lmo2t, rxr/Z177* and *sna*) are expressed by blastocoelar cells (a population of presumptive immunoctyes), whereas *six1/2* and *soxE* are expressed by pigment cells and coelomic pouch cells (probably prospective germ cells), respectively. In addition, two of the genes for which WMISH data are unavailable, *irf4* and *nfil3*, have vertebrate orthologs that play important roles in immune cell development, suggesting that these mRNAs might also be restricted to the blastocoelar cell lineage during normal development. We also examined the set of $\sim$100 non-DE effector genes that are upregulated both in Alx1 and Ets1 morphants and identified several proteins that are predicted to function in immune system development or physiology, including two Toll receptors (Sp-TlrL_9 and Sp-Tlr072) and a leukocyte receptor cluster member (Sp-Leng9L). Although further analysis of the expression patterns of these regulatory and effector genes in control embryos and morphants will be required, these findings support the view that a key function of *alx1* is to repress multiple, alternative mesodermal regulatory states, including the blastocoelar cell fate, in the LM progeny.

Surprisingly, Alx1 and Ets1 morphants exhibited a significant downregulation of *hox7*, a regulatory gene expressed in the aboral ectoderm, as well as *spec2c* and *spec2ce1-3*, two aboral ectoderm differentiation markers (supplementary material Tables S3 and S4). Two other aboral ectoderm regulatory genes, *dlx* and *msx*, were downregulated in Ets1 morphants. Our data therefore point to a previously unsuspected interaction between LM progeny and the aboral ectoderm that occurs before the early gastrula stage (i.e. the stage at which we analyzed gene expression).

2.4 Discussion

A complex sequence of PMC behaviors underlies the morphogenesis of the embryonic skeleton (Wilt and Ettensohn, 2007; Ettensohn, 2013). These behaviors require zygotic transcriptional inputs (Kurokawa et al., 1999; Ettensohn et al., 2003; Wu and McClay, 2007). Our work has provided the most complete picture to date of the effector genes that direct skeletogenesis and has revealed important features of the transcriptional control of these genes.

2.4.1 The identification of morphogenetic effector genes

The morphogenetic functions of some PMC effector genes are well understood. The spicule matrix proteins are a family of 15-20 closely related proteins occluded within the biomineral that influence its growth and physical properties, probably by regulating the conversion of amorphous calcium carbonate to the crystalline state (Wilt and Ettensohn, 2007; Gong et al., 2012; Rafiq et al., 2012). Non-fibrillar collagens produced by PMCs serve as an essential substrate for the cells (Wessel et al., 1991). Several PMC-specific, type I TM proteins, including P16, P58A and P58B, play essential roles in biomineral deposition (Cheers and Ettensohn, 2005; Adomako-Ankomah and Ettensohn, 2011). The precise biochemical functions of the P16 and P58 proteins are unknown, although P16 is phosphorylated and binds to hydroxyapatite (Alvares et al., 2009). A PMC-specific, GPI- anchored carbonic anhydrase is likely to be involved in biomineral remodeling (Livingston et al., 2006). All these proteins are associated with biomineralization, the major specialized function of the PMCs. Only one effector, VEGFR-Ig10, has been shown to mediate other aspects of the morphogenetic program of PMCs. This signaling receptor plays an essential role in PMC guidance; in addition, local VEGF signals acting through VEGFR-Ig10 control regional patterns of skeletal growth (Duloquin et al., 2007; Adomako-Ankomah and Ettensohn, 2013).

Our RNA-seq-based analysis has increased by approximately an order of magnitude the number of known PMC-enriched mRNAs and therefore provides a more complete picture of the output of this transcriptional network than was previously available. We identified a large number of putative effectors of skeletal morphogenesis that are candidates for further functional studies. In some cases (e.g. Fam20C and Otopetrin, see below), functions can be inferred from information concerning the vertebrate counterparts of these genes. Our work has also revealed specific proteins that are likely to account for pharmacological evidence that metalloproteases (Roe et al., 1989; Ingersoll and Wilt, 1998), calcium channels (Hwang and Lennarz, 1993) and ion transporters (Yasumasu et al., 1985; Mitsunaga et al., 1986; Fujino et al., 1987; Stumpp et al., 2012) are essential effectors of skeletogenesis.

2.4.2 Regulatory inputs into effector genes

Our findings have revealed important features of the regulatory inputs into the set of 420 effector genes. We focused on two key TFs in the PMC specification network: Ets1 and Alx1. These TFs provide regulatory inputs near the top of the regulatory network and are essential for PMC specification. Knockdown of Ets1 or Alx1 causes LM descendants to take on alternative mesodermal fates (Kurokawa et al., 1999; Ettensohn et al., 2003, 2007; Oliveri et al., 2008). Zygotic expression of both TFs is restricted to the LM lineage early in development; *alx1* transcription is activated selectively in LMs in the first interphase after these cells are born.

We found that of the 420 DE genes, more than half (223/420, 53%) received essential inputs from Ets1 and/or Alx1 (Fig. 2.5), the great majority of which ($\sim$90%) were positive. When only the most abundant mRNAs are considered, this value increased to 74% (i.e. 74/100 DE transcripts with the highest FPKM values in purified PMCs). We also noted that of the DE genes annotated with the GO terms 'biomineralization' or 'metalloprotease', 84% (32/38) were subject to regulatory inputs from one or both of these TFs. These findings demonstrate the central role of Ets1 and Alx1 in controlling the cell-specific identity of PMCs. At the same time, our analysis identified 197 DE genes that were not significantly affected by Ets1 or Alx1 knockdown. This number is likely to be inflated by the stringency of the Cuffdiff analysis; for example, many of these mRNAs showed modest changes in expression in morphants (e.g. 50-75% reduction in mRNA level) that were scored as non- significant. More importantly, we can assume that the MO knockdowns were incomplete. With these caveats in mind, we identified $\sim$150 DE mRNAs, many of which were very abundant, that showed changes in expression of $<$50% in both Ets1 and Alx1 morphants relative to controls. These findings indicate that Ets1/Alx1-independent circuits also make contributions to the specialized molecular properties of PMCs.

One of the most striking findings from this and previous work (Rafiq et al., 2012) is that many effector genes are regulated positively by both Ets1 and Alx1. Of the 223 DE genes with inputs from Ets1 and/or Alx1, $\sim$65% (144/223) were affected in both classes of morphants (Fig. 2.5). Several mechanisms might underlie this apparent co-regulation. First, Ets1 might regulate effector genes indirectly through its effect on Alx1 expression. Perturbation of Ets1 function does not affect the early phase of *alx1* expression, but suppresses the later phase (Oliveri et al., 2008; Sharma and Ettensohn, 2010). In our study, Ets1 knockdown reduced alx1 expression by 80%, whereas Alx1 knockdown had no effect on *ets1* expression. Moreover, of the 170 DE genes that were regulated by Ets1, 85% showed significant changes in expression following Alx1 knockdown.

Thus, most of the effects of Ets1 knockdown might be explained through the effect of Ets1 on Alx1 expression. It was reported previously that forced expression of Alx1 is unable to rescue the effects of Ets1 knockdown (Oliveri et al., 2008), which appears inconsistent with this model, but subsequent studies have shown that the effects of Alx1 are highly dosage dependent (Ettensohn et al., 2007; Damle and Davidson, 2011). Second, Ets1 and Alx1

might regulate effector genes in concert, via a feed-forward mechanism (e.g. Ets1>Alx1, Alx1>Effector X, Ets1>Effector X). Experimental evidence in support of this model has come from analysis of the cis-regulatory control of *cyp1*, which receives direct inputs from Dri (a target of *alx1*) and Ets1 (Amore and Davidson, 2006) and *sm50*, which receives direct inputs from Ets1 (Yajima et al., 2010). Third, Ets1 and Alx1 might regulate the expression of a common intermediary TF that provides essential inputs into many effector genes. If this is the case, then knockdowns of Ets1 and Alx1 would be expected to produce similar effects. Candidates include regulatory genes in the DE gene set that are downregulated both in Ets1 and Alx1 morphants (*alx4, cebpa, foxB* and *nk7*).

2.4.3 MAPK Signaling and the PMC GRN

The MAPK pathway plays a crucial role in PMC specification. Previous studies documented a transient, localized activation of ERK in the LM lineage shortly before ingression and showed that U0126, a selective MEK inhibitor, blocks PMC ingression and the expression of several terminal differentiation genes (Fernandez-Serra et al., 2004; Röttinger et al., 2004). The ERK/MAPK pathway is not required for the maternally driven activation of the network or the initial expression of early regulatory genes such as *alx1*, but becomes active at the blastula stage, when it functions to maintain the expression of *alx1* and possibly other regulatory genes (Fernandez-Serra et al., 2004; Sharma and Ettensohn, 2010). The activation of ERK in the LM-PMC lineage does not require signals from other cell populations (Fernandez-Serra et al., 2004; Röttinger et al., 2004). Significantly, Rottinger et al. (Röttinger et al., 2004) showed that Ets1 (which contains consensus MAPK phosphorylation and ERK docking sites) is a direct target of MEK/ERK signaling.

Our RNA-seq analysis showed that a surprisingly small fraction of the transcriptome is dependent upon MAPK signaling during early embryogenesis. We identified only 180 transcripts that exhibited significant changes in expression at 28-30 hpf in response to MEK inhibition. Strikingly, more than half of these transcripts (101/180) were contained in the DE collection. Our data are consistent with immunostaining studies indicating that ERK is selectively activated in the LM lineage and support the view that the PMC GRN is the principal target of MAPK signaling during early development. Later in gastrulation, p-ERK is also detected at the tip of the archenteron, where it plays a role in the specification of non-skeletogenic mesoderm (Fernandez-Serra et al., 2004; Röttinger et al., 2004). Genes that are sensitive to U0126 but not differentially expressed in PMCs (79/180 genes) might be direct targets of MAPK signaling in NSM cells or indirect targets in tissues that are dependent upon PMCs for their normal development. Of the 101 DE genes sensitive to U0126, most (74/101, 73%) were also found to be regulated by Ets1, strongly supporting the view that Ets1 is the key mediator of MAPK inputs into the PMC GRN. We also identified a small number of DE transcripts (17) that were sensitive to U0126 but not to knockdown of either Ets or Alx1; these included *tbr* and *mitf*, mRNAs that encode PMC-restricted TFs. The mechanism by which MAPK/ERK signaling regulates the

expression of these genes is unknown, although it is possible that the maternal pool of Ets1 protein (Yajima et al., 2010), which is not affected by MO knockdown and might be activated by MAPK, could be responsible. Our analysis also showed that >50% of the genes within the DE set that were sensitive to Ets1 knockdown were insensitive to U0126 (Fig. 2.5), suggesting that ERK-mediated phosphorylation is required for only a subset of the regulatory functions of Ets1. Lastly, our studies define a discrete, signal-dependent submodule of the larger genetic circuitry that controls PMC identity, represented by the subset ($\sim$1/4) of DE genes sensitive to MEK inhibition.

2.4.4 The evolution of biomineralization

The further elucidation of the genetic network that underlies skeletogenic specification and morphogenesis in echinoderms has important implications for reconstructing the evolution of biomineralization in metazoans. The fossil record documents a widespread and relatively synchronous emergence of biomineralization in many metazoan lineages during the Cambrian period (Knoll, 2003; Murdock and Donoghue, 2011). It is widely accepted that biomineralized structures, in the strictest sense, appeared independently in these lineages. For example, the first true mineralized vertebrate skeletons are thought to have appeared in ostracoderms, a group of stem gnathostomes, as a dermal skeleton, independently of the echinoderm skeleton (Donoghue and Sansom, 2002; Murdock and Donoghue, 2011). An important unanswered question, however, concerns the extent to which this occurred by exploiting a common 'toolkit', i.e. a set of ancestral biochemical and developmental pathways that was independently co-opted for biomineral formation in diverse animal taxa (Westbroek and Marin, 1998; Jackson et al., 2007; Murdock and Donoghue, 2011).

Our findings reveal new and surprising connections between genes that control biomineralization in modern echinoderms and vertebrates, despite the difference in biomineral content and micro- architecture in these taxa (Bottjer et al.). We found that PMCs selectively express the single sea urchin member of the otopetrin family. Otopetrin 1 is required for the formation of calcite otoliths/otoconia in vertebrates (Hurle et al., 2003; Hughes et al., 2004; Söllner et al., 2004). The precise biochemical function of this 12-pass TM protein is unknown, but it might play a role in regulating cytosolic Ca^{2+} levels in response to extracellular signals (Kim et al., 2010, 2011). We also identified in the DE gene set the *S. purpuratus* ortholog of Fam20C, an extracellular kinase that phosphorylates extracellular biomineralization proteins in vertebrates (Ishikawa et al., 2012; Tagliabracci et al., 2012), as a protein differentially expressed by PMCs. Other classes of proteins with conserved functions in biomineralization in echinoderms and vertebrates include collagens, matrix metalloproteases and carbonic anhydrases (Livingston et al., 2006; Krane and Inada, 2008; Wuthier and Lipscomb, 2011). Alx1 family members play conserved roles as upstream transcriptional regulators of skeletogenesis in both taxa (Ettensohn et al., 2003). Our studies therefore reveal an extensive, common biomineralization toolkit that was

likely to be present in the ancestral deuterostome and might have been exploited in diverse animal lineages.

2.5 Materials and Methods

Adult animals and embryo culture

Adult *Strongylocentrotus purpuratus* were obtained from Patrick Leahy (California Institute of Technology, USA). Spawning was induced by intracoelomic injection of 0.5 M KCl and embryos were cultured in artificial seawater (ASW) at 15°C in a temperature-controlled incubator.

PMC isolation

PMCs were isolated from early mesenchyme blastula stage embryos at 24 hours post-fertilization (hpf), as previously described (Harkey and Whiteley, 1980). Briefly, embryos were washed three times in calcium- and magnesium-free ASW (CMFSW), twice in 1 M glycine, and resuspended in bag isolation medium (per liter: 400 ml 1 M dextrose, 400 ml CMFSW, 200 ml distilled water). Embryos were dissociated by gentle pipetting. Basal lamina bags containing PMCs were collected using a sucrose step gradient. A 'non-PMC'(or 'other cell') fraction was collected from the same batch of embryos, also as described (Harkey and Whiteley, 1980). The same dissociation procedure was used except that embryos were washed only once in 1 M glycine to minimize rupturing of basal lamina bags. After resuspension in bag isolation medium, the sample was centrifuged at 650 g for 10 minutes, and the supernatant containing the non-PMC fraction was collected. The purity of isolated PMCs was >95% as assessed by immunostaining with monoclonal antibody 6a9 (Ettensohn and McClay, 1988). For analysis of transcript levels by RNA-seq, the PMC and non-PMC samples were isolated from two embryo cultures, derived from separate matings, which served as biological replicates.

Morpholino (MO) injections

MOs (Gene Tools) were injected into fertilized eggs as previously described (Cheers and Ettensohn, 2004), with the modification that eggs were fertilized in the presence of 0.1% (w/v) para-aminobenzoic acid to prevent hardening of the fertilization envelope. MO sequences (5'-3') were: SpAlx1, TATTGAGTTAAGTCTCGGCACGACA; SpEts1, GAACAGTGCATAGACGCCATGATTG; control, CCTCTTACCTCAGTTACAATTTATA.

The Ets1 and Alx1 MOs, both of which are translation-blocking, have been shown to be specific and effective (Ettensohn et al., 2003; Oliveri et al., 2008; Rafiq et al., 2012). MOs were injected at an initial concentration of 2 mM (Ets1) or 4 mM (Alx1). Injection solutions also contained 20% (v/v) glycerol and 0.16% (w/v) Texas Red dextran. The control MO was injected at the same concentration as the corresponding translation-blocking MO. For comparisons of transcript levels in controls and Ets1/Alx1 morphants by RNA-seq, 500 embryos were pooled at 28-30 hpf for each sample. Because there was some embryo-to-embryo variability in morphant phenotypes, we hand selected embryos that lacked PMCs at the start of invagination. A single embryo culture was used for the analysis of Alx1 knockdown and a separate culture for the analysis of Ets1 knockdown.

RNA-seq

Total RNA was extracted using the NucleoSpin RNA II Kit (Clontech) and precipitated with ethanol. For comparisons of transcript levels in PMCs and other cells, RT-PCR was performed using the RETROScript Kit (Clontech) and primers for several PMC-specific transcripts (*fc2*, *p133* and *can1*) and one housekeeping gene (*z12*), in order to confirm the expected difference in gene expression between the PMC and non-PMC samples. RNA samples were provided to the USC Epigenome Center and their quality was assessed using a BioAnalyzer. 600 ng-1g of total RNA was used for the construction of each Illumina HiSeq library. Sequencing was carried out with an Illumina HiSeq2000 machine (50 cycles, paired-end reads) with four to five indexed libraries in each lane. Approximately 40 million reads were obtained per sample. All data were analyzed using the open-source Tuxedo Suite (Langmead et al., 2009; Trapnell et al., 2012) with default parameters. TopHat (2.0.8b) was used to map sequence reads to the *S. purpuratus* transcriptome (Tu et al., 2012). The relative abundance of transcripts, represented by their FPKM (fragments per kilobase of transcript per million mapped reads) values, was estimated using Cufflinks (2.0.1). Cuffdiff (part of the Cufflinks package, (**?**)) was used to identify significant differences in the abundance of transcripts between samples (false discovery rate=0.05), and CummeRbund (2.0.0) (R/Bioconductor) was used for scatterplot analysis. All sequences were deposited in the NCBI Sequence Read Archive (SRA accession numbers SRP033427 and SRP031836).

Whole-mount *in situ* hybridization (WMISH)

Embryos were fixed for 1 hour at room temperature in 4% paraformaldehyde in ASW and stored at 4°C in 100% methanol. WMISH was carried out as previously described (Lepage et al., 1992; Duloquin et al., 2007)

Nanostring analysis

Quantitative analysis of transcript levels was carried out with a Nanostring nCounter system and a codeset corresponding to $\sim$90 genes in the PMC GRN, as previously described (Adomako-Ankomah and Ettensohn, 2013).

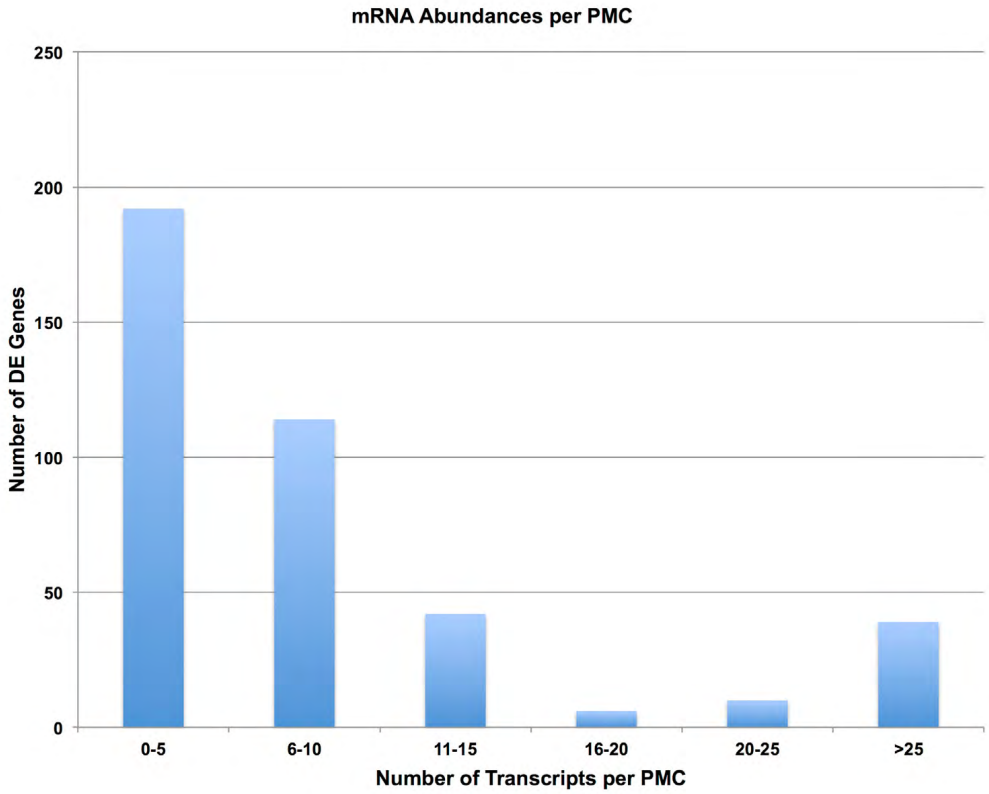
Analysis of temporal expression profiles

Temporal expression profiles of genes were obtained from the transcriptome data of Tu et al. (2012) and were based on raw FPKM values at 0, 10, 18, 24, 30, 40, 48, 56, 64 and 72 hpf. Temporal expression patterns were analyzed by hierarchical clustering using Euclidean distances (MATLAB, MathWorks).

U0126 treatment

Embryos were treated with 7 μM U0126 (Calbiochem) in DMSO continuously from the 2-cell stage and sibling control embryos were treated with DMSO alone. At 28-30 hpf, total RNA was collected for RNA-seq analysis. Several control and UO126-treated embryos were immunostained with monoclonal antibody 6a9 to confirm that PMC specification was blocked by U0126 treatment, as reported previously (Fernandez-Serra et al., 2004; Röttinger et al., 2004).

2.6 Supplementary Figures



Supplementary Figure 2.1: **Transcript abundances per PMC for genes differentially expressed in PMCs.** We first converted our FPKM values from control *S. purpuratus* embryos at 28-30 hpf to number of transcripts per embryo, using a conversion factor calculated from published data for 20 transcripts expressed at relatively constant levels during gastrulation (Materna et al., 2010). The number of transcripts per PMC was then determined, assuming 16 PMCs/embryo at the early gastrula stage and negligible expression in other cell types. The final formula applied was: Transcripts per PMC=(FPKM+1.3606)/0.08656.

Supplementary Figure 2.2: **Reproducibility of gene expression profling of Ets1 and Alx1 morphants.** To assess the reliability of our genome-wide identification of Ets1 and Alx1 targets, we compared the effects of Ets1 and Alx1 knockdowns as assessed by quantitative RNA-seq on the expression levels of 22 genes that had been shown to be regulated by these TFs in a previous study (Rafiq et al., 2012). Green: QPCR data from Rafiq et al. (2012). Red: Nanostring data from this study. Blue: RNA-seq data from this study. Vertical bars indicate the ratio of expression in morphant embryos relative to sibling controls at 28-30 hpf (equal expression=1 as indicated by the dotted lines; values <1 indicate reduced expression in morphant embryos). The effects of knockdowns on gene expression were highly reproducible across these experiments, which were carried out on embryos derived from three different male-female matings and each of which used a different method of transcript quantification.

Supplementary tables are available for download at http://dev.biologists.org/content/141/4/950.supplemental

Bibliography

Adams, D.K., Sewell, M.A., Angerer, R.C., and Angerer, L.M., 2011. Rapid adaptation to food availability by a dopamine-mediated morphogenetic response. *Nature Communications*, **2**:592.

Adomako-Ankomah, A. and Ettensohn, C.A., 2011. P58-A and P58-B: novel proteins that mediate skeletogenesis in the sea urchin embryo. *Developmental biology*, **353**(1):81–93. ISSN 1095-564X. doi:10.1016/j.ydbio.2011.02.021.

Adomako-Ankomah, A. and Ettensohn, C.A., 2013. Growth factor-mediated mesodermal cell guidance and skeletogenesis in the sea urchin embryo. *Development*, **140**:4212–4225.

Amore, G. and Davidson, E.H., 2006. cis-Regulatory control of cyclophilin, a member of the ETS-DRI skeletogenic gene battery in the sea urchin embryo. *Dev Biol*, **293**(2):555–64. doi:10.1016/j.ydbio.2006.02.024.

Bottjer, D.J., Davidson, E.H., Peterson, K.J., and Cameron, R.A., ????

Cameron, R.A., Samanta, M., Yuan, A., He, D., and Davidson, E., 2009. SpBase : the sea urchin genome database and web site. **37**(November 2008):750–754. doi:10.1093/nar/gkn887.

Cheers, M.S. and Ettensohn, C.A., 2004. Rapid microinjection of fertilized eggs. *Methods Cell Biol*, **74**:287–310.

Cheers, M.S. and Ettensohn, C.A., 2005. P16 is an essential regulator of skeletogenesis in the sea urchin embryo. *Developmental Biology*, **283**:384–396.

Croce, J.C. and McClay, D.R., 2010. Dynamics of Delta/Notch signaling on endomesoderm segregation in the sea urchin embryo. *Development*, **137**:83–91.

Damle, S. and Davidson, E.H., 2011. Precise cis-regulatory control of spatial and temporal expression of the alx-1 gene in the skeletogenic lineage of s. purpuratus. *Dev Biol*, **357**(2):505–17. doi:10.1016/j.ydbio.2011.06.016.

Davidson, E.H., 2010. Emerging properties of animal gene regulatory networks. *Nature*, **468**(7326):911–20. ISSN 1476-4687. doi:10.1038/nature09645.

Donoghue, P.C. and Sansom, I.J., 2002. Origin and early evolution of vertebrate skeletonization. *Microscopy Research and Technique*, **59**:352–372.

Duloquin, L., Lhomond, G., and Gache, C., 2007. Localized VEGF signaling from ectoderm to mesenchyme cells controls morphogenesis of the sea urchin embryo skeleton. *Development*, **134**(12):2293–302. doi:10.1242/dev.005108.

Emily-Fenouil, F., Ghiglione, C., Lhomond, G., Lepage, T., and Gache, C., 1998. GSK3beta/shaggy mediates patterning along the animal-vegetal axis of the sea urchin embryo. *Development*, **125**:2489–2498.

Ettensohn, C.A., 2006. The emergence of pattern in embryogenesis: regulation of beta-catenin localization during early sea urchin development. *Sci STKE*, **2006**(361):pe48. doi:10.1126/stke.3612006pe48.

Ettensohn, C.A., 2013. Encoding anatomy: Developmental gene regulatory networks and morphogenesis. *Wiley Periodicals*, **27**:1–27. ISSN 1526-968X. doi:10.1002/dvg.22380.

Ettensohn, C.A., Illies, M.R., Oliveri, P., and De Jong, D.L., 2003. Alx1, a member of the Cart1/Alx3/Alx4 subfamily of Paired-class homeodomain proteins, is an essential component of the gene network controlling skeletogenic fate specification in the sea urchin embryo. *Development*, **130**:2917–2928.

Ettensohn, C.A., Kitazawa, C., Cheers, M.S., Leonard, J.D., and Sharma, T., 2007. Gene regulatory networks and developmental plasticity in the early sea urchin embryo: alternative deployment of the skeletogenic gene regulatory network. *Development*, **134**(17):3077–87. doi:10.1242/dev.009092.

Ettensohn, C.A. and McClay, D.R., 1988. Cell lineage conversion in the sea urchin embryo. *Developmental Biology*, **125**:396–409.

Fernandez-Serra, M., Consales, C., Livigni, A., and Arnone, M.I., 2004. Role of the ERK-mediated signaling pathway in mesenchyme formation and differentiation in the sea urchin embryo. *Dev Biol*, **268**(2):384–402. doi:10.1016/j.ydbio.2003.12.029.

Flynn, C.J., Sharma, T., Ruffins, S.W., Guerra, S.L., Crowley, J.C., et al., 2011. High-resolution, three-dimensional mapping of gene expression using GeneExpressMap (GEM). *Dev Biol*, **357**(2):532–40. doi:10.1016/j.ydbio.2011.06.033.

Fuchikami, T., Mitsunaga-Nakatsubo, K., Amemiya, S., Hosomi, T., Watanabe, T., et al., 2002. T- brain homologue (HpTb) is involved in the archenteron induction signals of micromere descendant cells in the sea urchin embryo. *Development*, **129**:5205–5216.

Fujino, Y., Mitsunaga, K., and Yasumasu, I., 1987. nhibitory effects of omeprazole, a specific inhibitor of H+, K+-ATPase, on spicule formation in sea urchin embryos and in cultured micromere-derived cells. *Development, Growth and Differentiation*, **29**:591–597.

Gong, Y.U., Killian, C.E., Olson, I.C., Appathurai, N.P., Amasino, A.L., et al., 2012. Phase transitions in biogenic amorphous calcium carbonate. *Proceedings of the National Academy of Sciences of the United States of America*, **109**:6088–6093.

Guss, K.A. and Ettensohn, C.A., 1997. Skeletal morphogenesis in the sea urchin embryo: regulation of primary mesenchyme gene expression and skeletal rod growth by ectoderm-derived cues. *Development*, **124**:1899–1908.

Gustafson, T. and Wolpert, L.., 1967. Cellular movement and contact in sea urchin morphogenesis. *Biological Reviews of the Cambridge Philosophical Society*, **42**:442–498.

Harkey, M.A. and Whiteley, A.H., 1980. Isolation , Culture , and Differentiation

of Echinoid Primary Mesenchyme Cells. *Roux's Archives of Developmental Biology*, **122**(1896):111–122.

Hart, M.W. and Strathmann, R.R., 1994. Functional consequences of phenotypic plasticity in echinoid larvae. *Biological Bulletin*, **186**:291–299.

Hodor, P.G. and Ettensohn, C.A., 1998. The dynamics and regulation of mesenchymal cell fusion in the sea urchin embryo. *Developmental Biology*, **199**:111–124.

Hodor, P.G. and Ettensohn, C.A., 2008. Mesenchymal cell fusion in the sea urchin embryo. *Methods in Molecular Biology*, **475**:315–334.

Hodor, P.G., Illies, M.R., Broadley, S., and Ettensohn, C.A., 2000. Cell-substrate interactions during sea urchin gastrulation: migrating primary mesenchyme cells interact with and align extracellular matrix fibers that contain ECM3, a molecule with NG2-like and multiple calcium-binding domains. *Developmental Biology*, **222**:181–194.

Hughes, I., Blasiole, B., Huss, D., Warchol, M.E., Rath, N.P., et al., 2004. Otopetrin 1 is required for otolith formation in the zebrafish Danio rerio. *Developmental Biology*, **276**:391–402.

Hurle, B., Ignatova, E., Massironi, S.M., Mashimo, T., Rios, X., et al., 2003. Non-syndromic vestibular disorder with otoconial agenesis in tilted/mergulhador mice caused by mutations in otopetrin 1. *Human Molecular Genetics*, **12**:777–789.

Hwang, S.P. and Lennarz, W.J., 1993. Studies on the cellular pathway involved in assembly of the embryonic sea urchin spicule. *Experimental Cell Research*, **205**:383–387.

Illies, M.R., Peeler, M.T., Dechtiaruk, A.M., and Ettensohn, C.A., 2002. Identification and developmental expression of new biomineralization proteins in the sea urchin Strongylocentrotus purpuratus. *Development Genes and Evolution*, **212**:419–431.

Ingersoll, E.P. and Wilt, F.H., 1998. Matrix metalloproteinase inhibitors disrupt spicule formation by primary mesenchyme cells in the sea urchin embryo. *Developmental Biology*, **196**:95–106.

Ishikawa, H.O., Xu, A., Ogura, E., Manning, G., and Irvine, K.D., 2012. The Raine syndrome protein FAM20C is a Golgi kinase that phosphorylates bio-mineralization proteins. *PLoS ONE*, **7**:e42988.

Jackson, D.J., Macis, L., Reitner, J., Degnan, B.M., and Wörheide, G., 2007. Sponge paleogenomics reveals an ancient role for carbonic anhydrase in skeletogenesis. *Science*, **316**:1893–1895.

Kim, E., Hyrc, K.L., Speck, J., Salles, F.T., Lundberg, Y.W., et al., 2011. Missense mutations in Otopetrin 1 affect subcellular localization and inhibition of purinergic signaling in vestibular supporting cells. *Molecular and Cellular Neuroscience*, **46**:655–661.

Kim, T.K., Hemberg, M., Gray, J.M., Costa, A.M., Bear, D.M., et al., 2010. Widespread transcription at neuronal activity-regulated enhancers. *Nature*, **465**(7295):182–187. ISSN 0028-0836.

Kiyozumi, D., Takeichi, M., Nakano, I., Sato, Y., Fukuda, T., et al., 2012. Basement membrane assembly of the integrin 8 1 ligand nephronectin requires Fraser syndrome-associated proteins. *Journal of Cell Biology*, **197**:677–689.

Knoll, A.H., 2003. Biomineralization and evolutionary history. *Reviews in Mineralogy and Geochemistry*, **54**:329–356.

Krane, S.M. and Inada, M., 2008. Matrix metalloproteinases and bone. *Bone*, **43**:7–18.

Kurokawa, D., Kitajima, T., Mitsunaga-Nakatsubo, K., Amemiya, S., Shimada, H., et al., 1999. HpEts, an ets-related transcription factor implicated in primary mesenchyme cell differentiation in the sea urchin embryo. *Mechanisms of development*, **80**:41–52.

Langmead, B., Trapnell, C., Pop, M., and Salzberg, S.L., 2009. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome biology*, **10**(3):R25.1–R25.10. ISSN 1465-6914. doi:10.1186/gb-2009-10-3-r25.

Lepage, T., Ghiglione, C., and Gache, C., 1992. Spatial and temporal expression pattern during sea urchin embryogenesis of a gene coding for a protease homologous to the human protein BMP-1 and to the product of the Drosophila dorsal-ventral patterning gene tolloid. *Development*, **114**(1):147–63.

Livingston, B.T., Killian, C.E., Wilt, F., Cameron, A., Landrum, M.J., et al., 2006. A genome-wide analysis of biomineralization-related proteins in the sea urchin Strongylocentrotus purpuratus. *Developmental Biology*, **300**:335–348.

Logan, C.Y., Miller, J.R., Ferkowicz, M.J., and McClay, D.R., 1999. Nuclear beta-catenin is required to specify vegetal cell fates in the sea urchin embryo. *Development*, **126**(2):345–57.

Malinda, K.M., Fisher, G.W., and Ettensohn, C.A., 1995. Four-dimensional microscopic analysis of the filopodial behavior of primary mesenchyme cells during gastrulation in the sea urchin embryo. *Developmental Biology*, **172**:552–566.

Mann, K., Wilt, F.H., and Poustka, A.J., 2010. Proteomic analysis of sea urchin (Strongylocentrotus purpuratus) spicule matrix. *Proteome Science*, **8**:33.

Materna, S.C., Nam, J., and Davidson, E.H., 2010. High accuracy, high-resolution prevalence measurement for the majority of locally expressed regulatory genes in early sea urchin development. *Gene Expr Patterns*, **10**(4-5):177–84. doi:10.1016/j.gep.2010.04.002.

Miller, J., F.S.E. and McClay, D., 1995. Dynamics of thin filopodia during sea urchin gastrulation. *Development*, **121**:2501–2511.

Minokawa, T., Rast, J.P., Arenas-Mena, C., Franco, C.B., and Davidson, E.H., 2004. Expression patterns of four different regulatory genes that function during sea urchin development. *Gene expression patterns : GEP*, **4**(4):449–56. ISSN 1567-133X. doi: 10.1016/j.modgep.2004.01.009.

Mitsunaga, K., Fujino, Y., and Yasumasu, I., 1986. Change in the activity of Cl- ,HCO3(-)-ATPase in microsome fraction during early development of the sea urchin, Hemicentrotus pulcherrimus. *Journal of Biochemistry*, **100**:1607–1615.

Murdock, D.J.E. and Donoghue, P.C., 2011. Evolutionary origins of animal skeletal biomineralization. *Cells Tissues Organs*, **194**:98–102.

Oliveri, P., Tu, Q., and Davidson, E.H., 2008. Global regulatory logic for specification of an embryonic cell lineage. *Proceedings of the National Academy of Sciences of the United States of America*, **105**(16):5955–62. ISSN 1091-6490. doi:10.1073/pnas.0711220105.

Pennington, J.T. and Strathmann, R.R., 1990. Consequences of the calcite skeletons of planktonic echinoderm larvae for orientation, swimming, and shape. *Biological Bulletin*, **179**:121–133.

Peter, I.S. and Davidson, E.H., 2011. Evolution of gene regulatory networks controlling body plan development. *Cell*, **144**:970–985.

Peterson, R.E. and McClay, D.R., 2003. Primary mesenchyme cell patterning during the early stages following ingression. *Developmental Biology*, **254**:68–78.

Poustka, A.J., Kühn, A., Groth, D., Weise, V., Yaguchi, S., et al., 2007. A global view of gene expression in lithium and zinc treated sea urchin embryos: new components of gene regulatory networks. *Genome Biology*, **8**:R85.

Rafiq, K., Cheers, M.S., and Ettensohn, C.A., 2012. The genomic regulatory control of skeletal morphogenesis in the sea urchin. *Development*, **139**:579–590.

Ransick, A. and Davidson, E.H., 2006. cis-regulatory processing of Notch signaling input to the sea urchin glial cells missing gene during mesoderm specification. *Developmental Biology*, **297**:587–602.

Roe, J.L., Park, H.R., Strittmatter, W.J., and Lennarz, W.J., 1989. Inhibitors of metalloendoproteases block spiculogenesis in sea urchin primary mesenchyme cells. *Experimental Cell Research*, **181**:542–550.

Röttinger, E., Besnardeau, L., and Lepage, T., 2004. A Raf/MEK/ERK signaling pathway is required for development of the sea urchin embryo micromere lineage through phosphorylation of the transcription factor Ets. *Development*, **131**(5):1075–87. doi: 10.1242/dev.01000.

Sharma, T. and Ettensohn, C.A., 2010. Activation of the skeletogenic gene regulatory

network in the early sea urchin embryo. *Development*, **137**(7):1149–57. doi:10.1242/dev.048652.

Smyth, I. and Scambler, P., 2005. The genetics of Fraser syndrome and the blebs mouse mutants. *Human Molecular Genetics*, **14 Suppl.2**:R269–R274.

Sodergren, E., Weinstock, G.M., Davidson, E.H., Cameron, R.A., Gibbs, R.A., et al., 2006. The genome of the sea urchin Strongylocentrotus purpuratus. *Science*, **314**:941–952.

Söllner, C., Schwarz, H., Geisler, R., and Nicolson, T., 2004. Mutated otopetrin 1 affects the genesis of otoliths and the localization of Starmaker in zebrafish. *Development Genes and Evolution*, **214**:582–590.

Stathopoulos, A. and Levine, M., 2005. Genomic regulatory networks and animal development. *Developmental cell*, **9**:449–462.

Stumpp, M., Hu, M.Y., Melzner, F., Gutowska, M.A., Dorey, N., et al., 2012. Acidified seawater impacts sea urchin larvae pH regulatory systems relevant for calcification. *Proceedings of the National Academy of Sciences of the United States of America*, **109**:18192–18197.

Tagliabracci, V.S., Engel, J.L., Wen, J., Wiley, S.E., Worby, C.a., et al., 2012. Secreted kinase phosphorylates extracellular proteins that regulate biomineralization. *Science (New York, N.Y.)*, **336**(6085):1150–3. ISSN 1095-9203. doi:10.1126/science.1217817.

Trapnell, C., Roberts, A., Goff, L., Pertea, G., Kim, D., et al., 2012. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nature protocols*, **7**(3):562–78. ISSN 1750-2799. doi:10.1038/nprot.2012.016.

Tu, Q., Brown, C.T., Davidson, E.H., and Oliveri, P., 2006. Sea urchin Forkhead gene family: phylogeny and embryonic expression. *Developmental Biology*, **300**:49–62.

Tu, Q., Cameron, R.A., Worley, K.C., Gibbs, R.a., and Davidson, E.H., 2012. Gene structure in the sea urchin Strongylocentrotus purpuratus based on transcriptome analysis. *Genome research*, **22**(10):2079–87. ISSN 1549-5469. doi:10.1101/gr.139170.112.

Van Nostrand, E.L. and Kim, S.K., 2011. Seeing elegance in gene regulatory networks of the worm. *Current opinion in genetics & development*, **21**:776–786.

Verkman, A.S., 2011. Aquaporins at a glance. *Journal of Cell Science*, **124**:2107–2112.

Weitzel, H.E., Illies, M.R., Byrum, C.A., Xu, R., Wikramanayake, A.H., et al., 2004. Differential stability of beta-catenin along the animal-vegetal axis of the sea urchin embryo mediated by dishevelled. *Development*, **131**(12):2947–56. doi:10.1242/dev.01152.

Wessel, G.M., Etkin, M., and Benson, S., 1991. Primary mesenchyme cells of the sea urchin embryo require an autonomously produced, nonfibrillar collagen for spiculogenesis. *Developmental Biology*, **148**:261–272.

Westbroek, P. and Marin, F., 1998. A marriage of bone and nacre. *Nature*, **392**:861–862.

Wikramanayake, A.H., Huang, L., and Klein, W.H., 1998. beta-Catenin is essential for patterning the maternally specified animal-vegetal axis in the sea urchin embryo. *Proceedings of the National Academy of Sciences of the United States of America*, **95**:9343–9348.

Wilt, F.H. and Ettensohn, C.A., 2007. The morphogenesis and biomineralization of the sea urchin larval skeleton. *Handbook of Biomineralization: Biological Aspects and Structure Formation (ed. E. Bauerlein)*, pages 183–210.

Wu, S.Y. and McClay, D.R., 2007. The Snail repressor is required for PMC ingression in the sea urchin embryo. *Development (Cambridge, England)*, **134**(6):1061–70. ISSN 0950-1991. doi:10.1242/dev.02805.

Wunderlich, Z. and DePace, A.H., 2011. Modeling transcriptional networks in Drosophila development at multiple scales. *Current opinion in genetics & development*, **21**:711–718.

Wuthier, R.E. and Lipscomb, G.F., 2011. Matrix vesicles: structure, composition, formation and function in calcification. *Frontiers in Bioscience (Landmark Ed.)*, **16**:2812–2902.

Yajima, M., Umeda, R., Fuchikami, T., Kataoka, M., Sakamoto, N., et al., 2010. Implication of HpEts in gene regulatory networks responsible for specification of sea urchin skeletogenic primary mesenchyme cells. *Zoolog Sci*, **27**(8):638–46. doi:10.2108/zsj.27.638.

Yasumasu, I., Mitsunaga, K., and Fujino, Y., 1985. Mechanism for electrosilent Ca2+ transport to cause calcification of spicules in sea urchin embryos. *Experimental Cell Research*, **159**(80-90).

Chapter 3

Chromatin Accessibility Profiling Identifies *Cis*-regulatory Modules in an Early Embryonic Cell Lineage

This chapter consists of a submitted manuscript authored by Tanvi Shashikant, Jian Ming Khor and Charles A Ettensohn, currently in review at Genome Biology. I conducted all experiments and analyses described herein, other than some reporter assays conducted by Jian Ming Khor, some reporter constructs created by Dr. Debleena Dey and a few Python programs written by Siddharth Gurdasani. I co-wrote this manuscript with Dr. Charles A Ettensohn.

3.1 Abstract

Background: Gene regulatory networks (GRNs), which specify combinatorial interactions among regulatory genes and their inputs into downstream effectors, are powerful tools for delineating the genomic control of development. The GRN that underlies the development of skeletogenic cells (PMCs) in sea urchins is among the most comprehensive in any animal. While many gene interactions in this network have been described, only a handful of *cis*-regulatory modules (CRMs) have been identified. High-throughput discovery of PMC CRMs would significantly advance our understanding of this network.

Results: We used two independent methods of chromatin profiling, DNase-seq and ATAC-seq, to identify CRMs that control skeletogenic genes. 3,080 putative CRMs were identified, including 161 high-confidence CRMs pinpointed by both strategies. Putative PMC CRMs were preferentially located near skeletogenic genes and 29% of CRMs tested drove reporter gene expression specifically in PMCs. Consensus binding sites for two key skeletogenic transcription factors, Alx1 and Ets1, were enriched in these CRMs. CRMs as-

sociated with PMC effector genes were open in non-PMC lineages and most exhibited hypersensitivity by the 128-cell (late cleavage) stage.

Conclusions: Our work demonstrates the utility of differential chromatin accessibility for CRM discovery in embryonic tissues. The identification of hundreds of CRMs selectively active in PMCs will facilitate a comprehensive dissection of this important model developmental GRN and an improved understanding of GRN architecture more broadly. Our studies also reveal a surprising developmental history of a large cohort of cell type-specific CRMs, which are hypersensitive several hours before gene activation and are open in multiple embryonic lineages.

3.2 Background

The specification of cell fates during embryogenesis requires the precise spatiotemporal regulation of gene expression. A key regulatory step is gene transcription, which is mediated in large part by interactions between *cis*-regulatory modules (CRMs) and the transcription factors that bind to those modules. Recently, gene regulatory networks (GRNs) have been described that underlie cell type-specific transcriptional programs in developing embryos (Levine and Davidson, 2005; Peter and Davidson, 2015). GRNs are dynamic networks of regulatory genes (i.e., genes that encode transcription factors) and specify the combinatorial interactions among these genes as well as their inputs into downstream effectors. GRNs are proving to be powerful tools for understanding the genomic control and evolution of developmental processes (Peter and Davidson, 2016).

The sea urchin is a currently a preeminent model system for the elucidation of developmental GRNs. This is due to several advantageous characteristics of sea urchins for developmental studies (e.g., the ease with which many millions of synchronously developing embryos can be obtained) and to the development of experimental tools and genomic resources that support GRN analysis in sea urchins (Sea Urchin Genome Sequencing Consortium et al., 2006; Smith, 2008; Cameron et al., 2009; Tu et al., 2012; Cameron, 2014; Tu et al., 2014). Currently, the developmental GRNs that have been developed in sea urchins are among the most comprehensive for any animal.

The GRN that underlies the specification and morphogenesis of the skeleton-forming cells of the embryo (primary mesenchyme cells, or PMCs) is arguably the most complete GRN in sea urchins (Oliveri et al., 2008; Ettensohn, 2009; Rafiq et al., 2012; Ettensohn, 2013; Rafiq et al., 2014). The founder cells of the skeletogenic lineage are the four large micromeres that arise at the vegetal pole of the cleavage stage embryo. During gastrulation, the descendants of the large micromeres ingress into the blastocoel, migrate directionally, and fuse to form a syncytial network within which they deposit the calcified biomineral that serves as the embryonic skeleton (Wilt and Ettensohn, 2007). The GRN deployed in this lineage is activated by maternal factors localized at the veg-

etal pole of the egg (Logan et al., 1999; Weitzel et al., 2004). These maternal factors act through a double-repression system involving the transcriptional repressor, *pmar1/micro1* (Oliveri et al., 2002; Yamazaki et al., 2005) to activate a small set of early zygotic regulatory genes, including *Sp-ets1* (Kurokawa et al., 1999) and *Sp-alx1* (Ettensohn et al., 2003), selectively in the large micromere-PMC lineage. The products of these genes engage additional layers of regulatory genes and interactions among the various regulatory genes stabilize the transcriptional network and drive it forward (Oliveri et al., 2008). Recently, transcriptome profiling has been used to identify hundreds of downstream effector genes in the PMC GRN (Rafiq et al., 2014; Barsi et al., 2014). These effectors regulate critically important aspects of skeletal morphogenesis, including PMC migration, PMC-PMC fusion, and biomineral formation (Peled-Kamar et al., 2002; Duloquin et al., 2007; Adomako-Ankomah and Ettensohn, 2011, 2013; Ettensohn and Dey, 2017; Sun and Ettensohn, 2017)

Currently, the GRNs that have been constructed for sea urchins, including the PMC GRN, comprise positive and negative regulatory interactions that have been revealed by perturbing the function of specific regulatory genes and measuring effects on the expression of other genes in the network. Thus, they are maps of functional (epistatic) interactions that, in most cases, do not discriminate between direct and indirect interactions. For a relatively small number of genes, detailed mutational studies of CRMs have been carried out using reporter constructs and direct transcriptional inputs have been identified. For example, with respect to the PMC GRN, CRMs of *Sp-sm50* (Makabe et al., 1995), *Sp-sm30* (Akasaka et al., 1994; Yamasu and Wilt, 1999), *Sp-tbr* (Wahl et al., 2009) and *Sp-alx1* (Damle and Davidson, 2011) have been experimentally dissected to varying extents.

A major roadblock to a more detailed understanding of the architecture of this (and other) developmental GRNs is the challenge of identifying relevant CRMs. Experimental analysis of CRMs is currently the gold-standard for elucidating direct interactions between specific regulators and their target genes (Peter and Davidson, 2015). Therefore, the high-throughput identification of CRMs is a critically important step in developing comprehensive GRNs. Evolutionary conservation has been used to assist in the identification of putative CRMs (Yuh et al., 2002), but by itself this approach is less than satisfactory. Methods have also been developed that allow multiplexing of barcoded reporters to enhance *cis*-regulatory analysis (Nam and Davidson, 2012), but these are technically challenging and would benefit from accurate, high-throughput methods for CRM identification.

Genome-wide techniques such as DNase-seq and ATAC-seq have been used to identify regions of open chromatin in a variety of cell types (Crawford et al., 2006; John et al., 2013; Buenrostro et al., 2015). These methods rely on the local depletion of nucleosomes at promoters and CRMs that renders these regions hypersensitive to enzymes such as DNase I and Tn5 transposase. Hypersensitive DNA fragments are selectively isolated, sequenced, and mapped to the genome. Several studies have shown that cultured cell lines and adult tissues have patterns of chromatin accessibility characteristic of those cell types (Song et al., 2011; Ernst et al., 2011; Natarajan et al., 2012; Thurman et al., 2012).

In a few cases, cell type-specific patterns of chromatin accessibility have been used as a primary criterion for CRM discovery (Xiong et al., 2013; Wilken et al., 2015; Pearson et al., 2016).

In order to obtain a comprehensive understanding of the gene regulatory program deployed in the large micromere-PMC lineage, we set out to identify relevant functional CRMs in a high-throughput manner and reveal potential regulatory inputs. We used a combination of DNase-seq and ATAC-seq to identify a high-confidence set of CRMs that regulate gene expression in the skeletogenic lineage and showed that a large fraction of these CRMs drive PMC-specific expression of GFP reporter plasmids. Our work demonstrates the value of using differential chromatin accessibility for the high-throughput identification of CRMs in early embryonic cells. Furthermore, our identification of hundreds of CRMs selectively active in PMCs will facilitate a comprehensive dissection of this important model developmental GRN and an improved understanding of GRN architecture more broadly. Our studies also reveal a surprising developmental history of a large suite of lineage-specific CRMs, which we find are hypersensitive several hours before cell type-specific transcripts are expressed and are open in multiple embryonic cell lineages. The latter may reflect the pluripotency of early sea urchin embryonic cells or the association of these CRMs with repressors in non-skeletogenic cells.

3.3 Results

3.3.1 Analysis of U0126-dependent, hypersensitive sites identified by DNase-seq

We used DNase-seq in combination with pharmacological ablation of PMCs to identify candidate PMC CRMs. DNase-seq was performed on three biological replicates from separate matings. Each replicate consisted of two samples: control mesenchyme blastula stage embryos and sibling U0126-treated, PMC(-) embryos (Fig. 3.1A). A total of six Illumina libraries were generated and sequenced. Sequence reads were analyzed using a bioinformatics pipeline consisting of various open-source tools and two custom Python programs (Fig. 3.1B). We found that the peaks identified in replicate samples were highly concordant, with an average pairwise Pearson's correlation coefficient of 0.955 (Supp. Fig. 3.1A). An average of 23.5 million 50 bp single-end reads were obtained per sample, of which 19 million reads (80.8%) on average were mapped to the *S. purpuratus* genome (Table 3.1). After PCR duplicate removal and read count equalization, an average of 256,007 peaks (average size = 442 bp) were called from the resulting 14.7 million mapped reads per sample. The average fraction of reads in peaks was 0.55 and the Reference Peak Set (RPS) consisted of 157,108 peaks of average size 637 bp. The RPS covered 10.68% of the genome. 1,659 peaks were identified that had significantly elevated signal (nominal p-value <0.1) in control embryos compared to PMC(-) embryos. We refer to such peaks

as "DNase-seq differential peaks" (see Fig. 3.1E for an example and Supp. Table 1 for coordinates of all DNase-seq differential peaks identified). We found that 258 DNase-seq differential peaks were within 10 kb of genes differentially expressed by PMCs ("PMC DE genes") (Rafiq et al., 2014). Differential peaks were much more likely to be located within 10 kb of DE genes than non-differential peaks; this difference was significant by Fisher's exact test (p-value <2.2e-16; 5.31-fold enrichment).

DNase-seq Sequencing Information	
Avg. no. sequenced reads/sample	23.5 M
Avg. no. reads mapped/sample	19 M (80.8%)
Avg. no. reads mapped/sample after duplicate removal and equalization	14.7 M (62.5%)
Avg. no. peaks called/sample	256,007
Avg. fraction reads in peaks	0.55
No. peaks in RPS	157,108
Avg. size of peak in RPS	637 bp
Genome coverage of RPS	10.68%
No. differential peaks (p < 0.1)	1,659
No. differential peaks within 10kb of a gene	1,287
No. differential peaks within 10 kb of a PMC DE gene	258; 5.31-fold enrichment p < 2.2e-16

Table 3.1: Sequence analysis information for DNase-seq samples. A reference peak set (RPS) was generated by first merging all highly concordant peaks among control embryo and PMC(-) embryo replicates separately, and then merging these two peak sets. 1,659 peaks with with nominal p-values under 0.1 (calculated by DESeq2) were determined to be enriched in the control embryos compared to PMC(-) embryos: these are DNase-seq differential peaks. 258 of these peaks were found to be within 10 kb of PMC DE genes, a highly significant enrichment (5.31-fold enrichment; p <2.2e-16) as determined by Fisher's exact test.

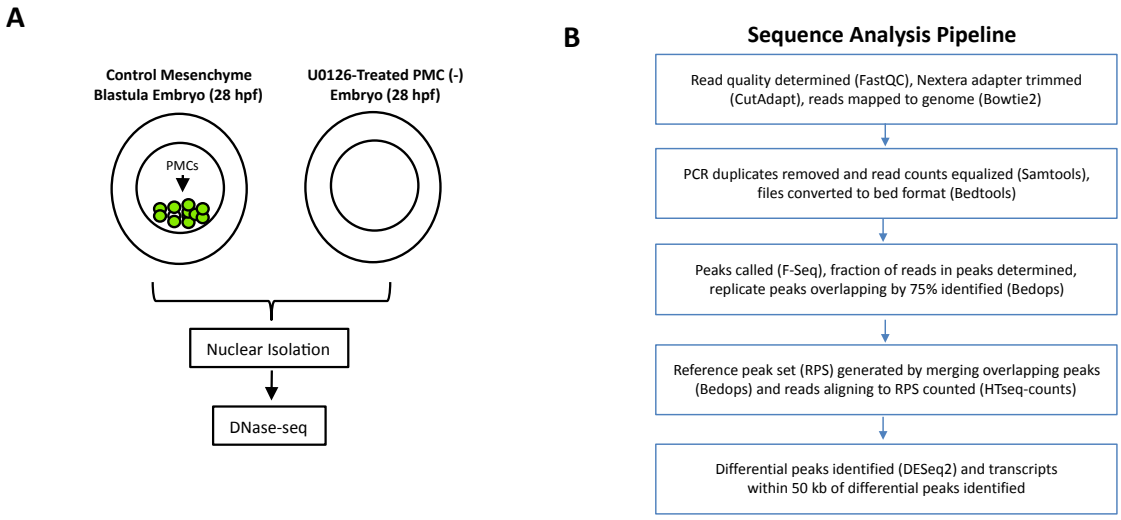
We mapped the location of each peak in the RPS relative to the nearest gene. Peak locations were classified as follows: Upstream (5′): The 3′ end of the peak was within 1-10 kb upstream of the 5′ end of the first exon; Promoter: The 3′ end of the peak was within 1 kb upstream of the 5′ end of the first exon; Within Gene Body: The 5′ end of the peak was within introns or exons; Downstream (3′): The 5′ end of the peak was within 10 kb downstream of, and did not overlap, the 3′ end of the last exon; Distal: No portion of the peak was within 10 kb of a gene. We found that 42% of the peaks in the RPS were distal, 30% were within gene bodies, 12% were downstream of genes, 9% were upstream of genes and 7% were closely associated with putative promoter regions (Fig. 3.1C). Of the peaks

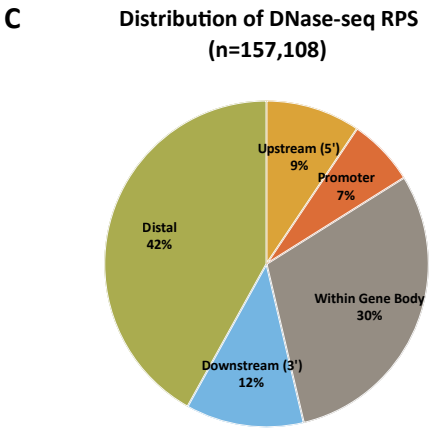
found within gene bodies, the majority (∼90%) were in introns. Of the 1,659 DNase-seq differential peaks, 22% were distal, 47% were within gene bodies, 9% were downstream of genes, 9% were upstream of genes and 13% were closely associated with putative promoter regions (Fig. 3.1D). Of the peaks found within the gene body, the majority (∼90%) were in introns. These data revealed a significant enrichment of DNase-seq differential peaks in promoter regions (Fisher's exact p-value = 1.10e-15; 1.97-fold enrichment) and within gene bodies (Fisher's exact p-value = 1.10e-15; 1.55-fold enrichment) and a significant depletion of distal peaks (Fisher's exact p-value = 1.10e-15; 1.32-fold depletion) relative to the RPS. Thus, the putative PMC CRMs identified by our DNase-seq analysis were more likely to be found close to PMC DE genes and more likely to be located within gene bodies and in putative promoter regions than was true of DNase-seq peaks as a whole.

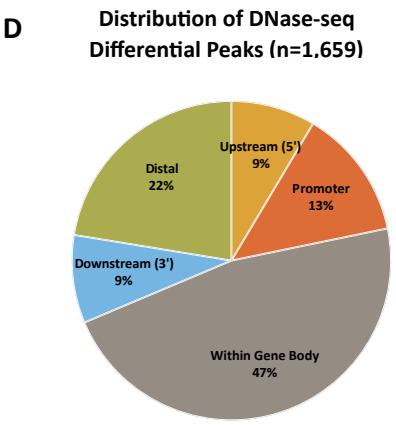
Of the 1,659 DNase-seq differential peaks, 1,287 peaks were within 10 kb of 1,216 genes in the *S. purpuratus* genome. Of these 1,216 genes, 400 have been assigned to functional (GO) categories, as annotated in Echinobase (Sea Urchin Genome Sequencing Consortium et al., 2006; Cameron et al., 2009; Tu et al., 2012). Biomineralization and transcription factor functional categories were highly enriched (adjusted Fisher's p-value <8e-07, 3.46-fold avg. enrichment) in the set of genes that were within 10 kb of DNase-seq differential peaks (see Supplementary Fig. 3.2A). The overrepresentation of biomineralization genes was striking as the primary biological function of PMCs is to secrete the calcified endoskeleton and biomineralization gene constitute the largest functional class of PMC DE genes (Rafiq et al., 2014).

U0126 has been shown to selectively block PMC specification (Fernandez-Serra et al., 2004; Röttinger et al., 2004). Nevertheless, because U0126 inhibited MAPK signaling throughout the developing embryo and blocked PMC specification at an early developmental stage, we considered it likely that the 1,659 DNase-seq differential peaks represented not only PMC CRMs but CRMs active in other tissues and sensitive either to MAPK signaling or to cell signals ordinarily provided by PMCs. For these reasons, PMC CRMs are likely to represent a subset of the 1,659 DNase-seq differential peaks. As DNase-seq requires at least 10 million nuclei as starting material, it was not feasible to perform DNase-seq using nuclei extracted from isolated PMCs, which are difficult to obtain in such high numbers. Therefore, to enhance the specificity of PMC CRM detection, we used ATAC-seq, a method that requires relatively few nuclei, to compare the chromatin accessibility patterns of isolated PMCs and other (non-PMC) cells at the mesenchyme blastula stage. We reasoned that by correlating data obtained from two independent methods of chromatin accessibility mapping we could identify a set of high-confidence CRMs that mediate PMC gene expression.

Figure 3.1: DNase-seq Sample Preparation and Sequence Analysis







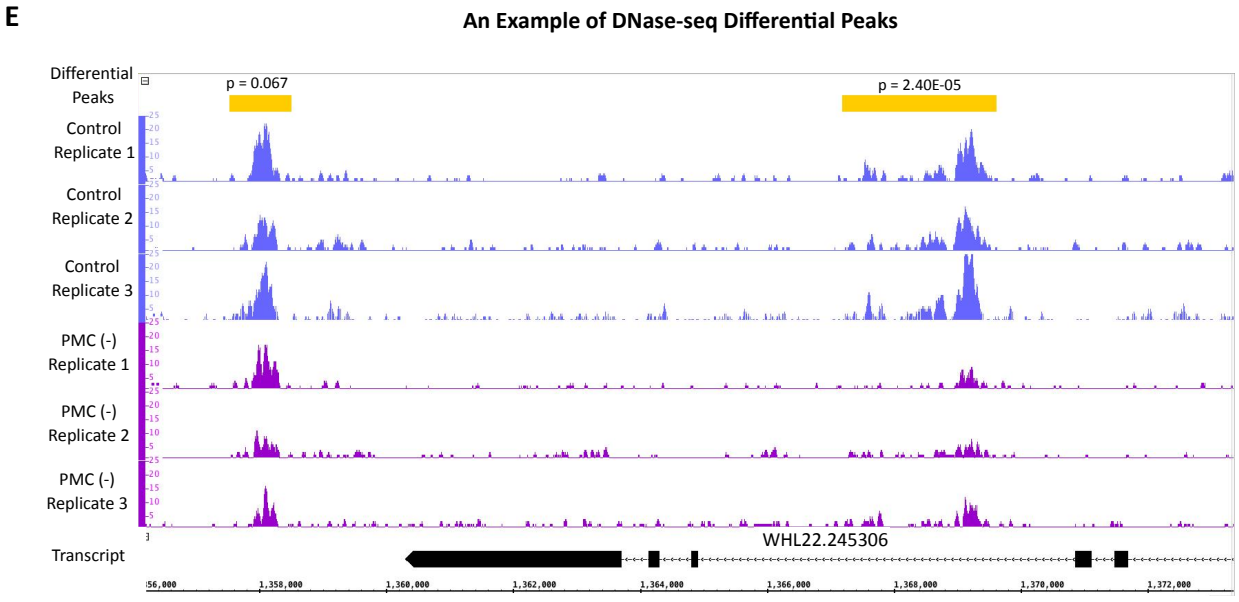


Figure 3.1: DNase-seq Sample Preparation and Sequence Analysis. A) *S. purpuratus* embryos were treated with U0126 at the 2-cell stage to obtain PMC(-) embryos. Control and U0126-treated embryos were cultured for 28 hours at 15°C in triplicate using three pairs of male and female sea urchins. Nuclei were isolated and DNase-seq was carried out, followed by Illumina sequencing. B) An outline of the bioinformatics pipeline used for DNase-seq and ATAC-seq sequence analysis. C) Distribution of DNase-seq peaks in the RPS with respect to the closest gene. See methods for definitions of peak locations. D) Distribution of DNase-seq differential peaks with respect to the closest gene. Compared to the distribution of all peaks in the RPS (see Fig 3.1C), there is a significant enrichment of peaks in putative promoter regions and within the gene body, and a significant depletion of peaks in distal regions. E) An example of DNase-seq differential peaks. The differential peaks (yellow rectangles) are located near the WHL22.245306 transcript. The aligned reads for each replicate are visualized as traces, and the differences in peak magnitude are clear when comparing control whole embryos (violet peak trace) to PMC(-) embryos (dark purple trace). Nominal p-values for differential peaks are indicated.

3.3.2 Analysis of differentially hypersensitive sites in PMCs identified by ATAC-seq

ATAC-seq was performed on three biological replicates from separate matings. Each replicate consisted of two samples: isolated PMCs and all other (non-PMC) cells obtained from mesenchyme blastula embryos at 28 hpf (Fig. 3.2A). A total of six Illumina libraries were generated and sequenced. An average of 89 million 76 bp single-end reads were obtained per sample, of which 69 million reads (77.5%) on average mapped to the *S. purpuratus* genome. After PCR duplicate removal and read count equalization, an average of 367,113 peaks were called from the resulting 43 million mapped reads per sample. Two sets of replicates were highly concordant, with an average pairwise Pearson's correlation coefficient of 0.915 (Supp. Fig. 3.1B). One replicate, however, was less concordant (Pearson's correlation coefficient threshold $\leq$0.8) and was therefore not included in the analysis. The average fraction of reads in peaks was 0.635 and the ATAC-seq RPS consisted of 295,441 peaks (average size = 597 bp) (Table 3.2). The RPS covered 18.84% of the genome. 1,582 peaks were identified that had significantly elevated signal (nominal p-value $<$0.2) in isolated PMCs compared with the non-PMC ("Other Cell") cell fraction. We refer to these peaks as "ATAC-seq differential peaks" (see Fig. 3.2D for examples and Supp. Table 2 for the coordinates of all ATAC-seq differential peaks identified). We found that 275 ATAC-seq differential peaks were within 10 kb of PMC DE genes. Differential peaks were much more likely to be located within 10 kb of DE genes than non-differential peaks; this difference was significant by Fisher's exact test (p-value $<$2.2e-16; 5.99-fold enrichment).

ATAC-seq Sequencing Information	
Avg. no. sequence reads/sample	89 M
Avg. no. reads mapped/sample	69 M (77.5%)
Avg. no. reads/sample after duplicate removal and equalization	43 M (48.3%)
Avg. no. peaks called/sample	367,113
Avg. fraction reads in peaks	0.635
No. peaks in RPS	295,441
Avg. size of peak in RPS	597 bp
Genome coverage of RPS	18.84%
No. differential peaks ($p < 0.2$)	1,582
No. differential peaks within 10 kb of genes	1,063
No. differential peaks within 10 kb of PMC DE genes	275; 5.99-fold enrichment, $p < 2.2e-16$

Table 3.2: Sequence analysis information for ATAC-seq samples. A reference peak set (RPS) was generated by first merging all highly concordant peaks among isolated PMC and other PMC(-) cell replicates separately, and then merging these two peak sets. 1,582 peaks with nominal p-values under 0.2 (calculated by DESeq2 (Love et al., 2014)) were determined to be be enriched in the isolated PMCs compared to the other non-PMC cells: these are ATAC-seq differential peaks. 275 of the 1,582 differential peaks were found to be within 10 kb of PMC DE genes, a highly significant enrichment ($p < 2.2e-16$; 5.99-fold enrichment) as determined by Fisher's exact test.

Genes closest to the peaks in the RPS were identified and the location of each peak with respect to the closest gene was determined. Peak locations were defined as for DNase-seq analysis (see above). Of all the ATAC-seq peaks in the RPS, 40% of the peaks were distal, 37% were within gene bodies, 11% were downstream of genes, 9% were upstream of genes and 5% were closely associated with putative promoter regions (Fig. 3.2B). Of the 1,582 ATAC-seq differential peaks, 33% of the peaks were distal, 47% were within gene bodies, 9% were downstream of genes, 7% were upstream of genes and 4% were closely associated with putative promoter regions (Fig. 3.2C). Thus, as in the case of DNase-seq differential peaks, we observed an enrichment of ATAC-seq differential peaks within gene bodies (Fisher's exact p-value = 1.10e-15; 1.29-fold enrichment) and a depletion in distal regions (Fisher's exact p-value = 9.25e-08; 1.21-fold depletion) compared to the RPS.

Figure 3.2: ATAC-seq Sample Preparation and Sequence Analysis

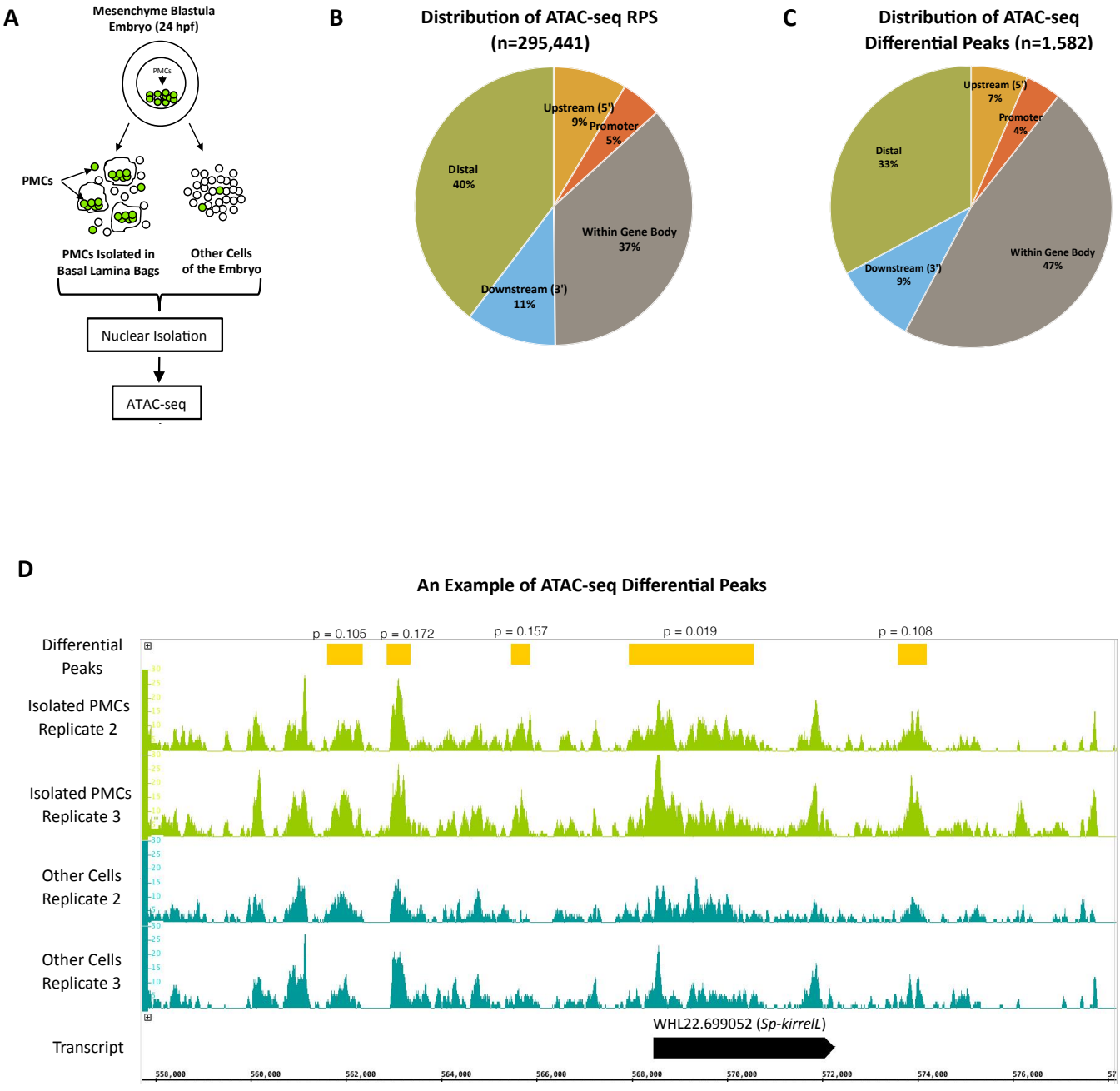


Figure 3.2: ATAC-seq Sample Preparation and Sequence Analysis A) *S. purpuratus* embryos were cultured for 24 hours at 15°C in triplicate using three pairs of male and female sea urchins. PMCs and other cells were isolated (Harkey and Whiteley, 1980) and ATAC-seq libraries were created and sequenced. Sequence reads were analyzed by the bioinformatics pipeline described in Fig. 3.1B. Peaks called in one sample replicate did not pass our Pearson's correlation threshold and were not analyzed further. B) The distribution of ATAC-seq peaks in the RPS with respect to the closest gene. Peak locations are defined in Fig. 3.1C. C) The distribution of ATAC-seq differential peaks with respect to the closest gene. In comparison to the distribution of all peaks in the RPS (see Fig 3.2B), there is a significant enrichment of peaks within the gene body and a significant depletion of peaks in distal regions. D) An example of ATAC-seq differential peaks. The differential peaks (yellow rectangles) are located near the *Sp-kirrelL* gene, a PMC DE gene. The aligned reads for each replicate are visualized, and the difference in peak magnitudes can be seen when comparing differential peaks in the isolated PMC replicates (light green peak trace) to the other cell replicates (dark green trace). Nominal p-values for differential peaks are indicated.

Of the 1,582 ATAC-seq differential peaks, 1,063 peaks were within 10 kb of 1,110 genes in the *S. purpuratus* genome. Of these 1,110 genes, 326 have been assigned to functional (GO) categories, as annotated in Echinobase (Sea Urchin Genome Sequencing Consortium et al., 2006; Cameron et al., 2009; Tu et al., 2012). Biomineralization and metalloprotease functional categories were highly enriched (adjusted Fisher's exact test p-value $<$0.05; 3.26-fold avg. enrichment) in genes that were within 10 kb of ATAC-seq differential peaks (see Supplementary Fig. 3.2B). As noted above, biomineralization is the principal biological function of PMCs and biomineralization gene constitute the largest functional class of PMC DE genes. In addition, pharmacological studies have shown that metalloprotreases play a critically important role in skeletogenesis (Roe et al., 1989; Ingersoll and Wilt, 1998).

3.3.3 Correspondence between DNase-seq and ATAC-seq datasets

We examined the extent to which the data obtained by these two independent chromatin accessibility mapping methods were congruent. First, we assessed the general correspondence between the genome-wide chromatin accessibility profiles obtained by the two methods (Table 3.3). The total number of peaks in the ATAC-seq RPS was almost twice that of the DNase-seq RPS (295,441 and 157,108 peaks, respectively), although the average peak sizes of the two RPSs were very similar (597 and 637 bp, respectively). The larger number of called peaks in the ATAC-seq RPS may have been due to the greater depth of sequencing and/or to a lower level of noise in these data (we found that the average FRIP score was slightly higher in the ATAC-seq data). Despite these differences, when we

compared the RPSs derived from DNase-seq and ATAC-seq data, we observed a high degree of correspondence. A very large fraction of the DNase-seq peaks (88%) overlapped ATAC-seq peaks by at least 1 nt. The fraction of the larger ATAC-seq RPS that overlapped DNase-seq peaks by at least 1 nt was, of course, smaller (44%) as there were many more peaks in the ATAC-seq RPS.

Correspondence Between ATAC-seq and DNase-seq Datasets	
No. peaks in DNase-seq RPS	157,108
No. peaks in ATAC-seq RPS	295,441
No. of DNase-seq RPS peaks overlapping ATAC-seq RPS peaks	138,145 (88% of DNase-seq RPS)
No. of ATAC-seq RPS peaks overlapping DNase-seq RPS peaks	130,552 (44% of ATAC-seq RPS)
No. of DNase-seq differential peaks (p < 0.1)	1,659 peaks
No. of ATAC-seq differential peaks (p < 0.2)	1,582 peaks
No. of overlapping differential peaks	161 (p = 2.2e-16)
No. of overlapping differential peaks within 10 kb of PMC DE genes	73 (p < 5.5e-10); 2.38-fold avg. enrichment

Table 3.3: Correspondence between ATAC-seq and DNase-seq datasets. 88% of all DNase-seq peaks in the RPS overlap with ATAC-seq peaks in the RPS by at least 1 nt, while 44% of the ATAC-seq RPS overlaps with the DNase-seq RPS by at least 1 nt. 161 peaks are present in both the DNase-seq and ATAC-seq differential peak set, overlapping by at least 75% in one direction. This overlap is highly significant (p <2.2e-16) as determined by Fisher's exact test. Of these 161 overlapping peaks, 73 are within 10 kb of PMC DE genes, a highly significant enrichment (Fisher's exact test p <5.5e-10; 2.38-fold avg. enrichment) compared to the enrichment of PMC DE genes observed with the ATAC-seq and DNase-seq differential peaks alone.

We next examined the extent of overlap between the 1,659 DNase-seq differential peaks and the 1,582 ATAC-seq differential peaks (Table 3.3). All DNase-seq differential peaks with 75% or more of their sequence overlapping one or more ATAC-seq differential peaks were merged with all ATAC-seq differential peaks that had 75% or more of their sequence overlapping one or more DNase-seq differential peaks. This operation generated a new set of 161 peaks common to the DNase-seq and ATAC-seq datasets: we call these peaks "overlapping differential peaks" (see Fig. 3.3A for examples of overlapping peaks and and Supp. Table 5 for the coordinates of all 161 merged, overlapping peaks). Although the number of overlapping differential peaks was not large, the probability that the observed degree of overlap between the DNase-seq differential peaks and the ATAC-seq differential peaks occurred by chance was vanishingly small (p-value <2.2e-16 by Fisher's exact test), demonstrating that the two independent datasets indeed converged on related populations of differential peaks. The degree of overlap may have been reduced by

a combination of factors, including possible effects of U0126 on tissues other than PMCs, the small amount of contamination by other cell types in the PMC preparations, differences in the sensitivities of the two techniques, or other unknown factors.

A large fraction of peaks in the overlapping differential peaks were within 10 kb of a PMC DE gene (73/161 peaks; 45%). This represented a significant enrichment compared to that observed in the ATAC-seq and DNase-seq differential peaks as a whole (Fisher's exact test p-value <5.5e-10; 2.38-fold enrichment). Of the 161 overlapping differential peaks, 136 peaks were within 10 kb of 135 genes in the *S. purpuratus* genome. Of these 135 genes, 55 have been assigned to functional (GO) categories, as annotated in Echinobase (Sea Urchin Genome Sequencing Consortium et al., 2006; Cameron et al., 2009; Tu et al., 2012). The biomineralization functional category was very highly enriched (adjusted Fisher's exact p-value = 2.73e-12; 19.61-fold enrichment) in genes that were within 10 kb of overlapping differential peaks (see Supplementary Fig. 3.2C), further supporting the view that the CRMs associated with these genes are active in PMCs.

In a previous study (Rafiq et al., 2014), the expression patterns of 420 PMC-enriched transcripts were classified into four clusters based on the developmental transcriptome data of Tu and co-workers (Tu et al., 2012). (Fig 3.3B). Cluster 1 consisted of 104 transcripts with maximal expression between 0-10 hpf, cluster 2 consisted of 136 transcripts with maximal expression between 40-72 hpf, cluster 3 consisted of 155 transcripts with maximal expression between 24-40 hours hpf, and cluster 4 consisted of 25 transcripts with maximal expression between 18-24 hpf. When we assigned the 62 PMC-enriched transcripts located within 10 kb of overlapping differential peaks to the above clusters, we observed a significant enrichment of these transcripts in Cluster 3 (Fisher's exact test p-value = 0.0173) (Fig. 3.3C). Cluster 3 genes were expressed maximally at a time that corresponded closely to the developmental stage we used for chromatin accessibility profiling and included a disproportionate number of genes with roles in skeletal development. A corresponding reduction in the proportions of transcripts in clusters 1 and 2 was also observed, but these differences were not statistically significant.

We also binned the 420 PMC DE genes into four classes based on their expression levels in PMCs at 24 hpf, using the RNA-seq data of Rafiq et al. (Rafiq et al., 2014). The "high expression" class (70 genes) had expression levels between 2512-100 FPKM, the "medium expression" class (117 genes) had expression levels between 99-40 FPKM, the "low expression" class (127 genes) had expression levels between 39-15 FPKM, and the "very low expression" class (106 genes) had expression levels between 14-0 FPKM. We found that the set of PMC DE genes that were within 10 kb of all differential peak sets showed a different distribution of expression levels than PMC DE genes as a whole. Specifically, genes near overlapping, differential peaks were significantly more likely to be in the "high expression" class (Fisher's exact p-value = 0.00042) and less likely to be in the "very low expression" class (Fig. 3.3D). This finding was consistent with our observation that overlapping, differential peaks tended to lie near biomineralization genes, most of which are expressed at high levels in PMCs at this stage (Rafiq et al., 2014).

Figure 3.3: Correspondence Between DNase-seq and ATAC-seq Datasets

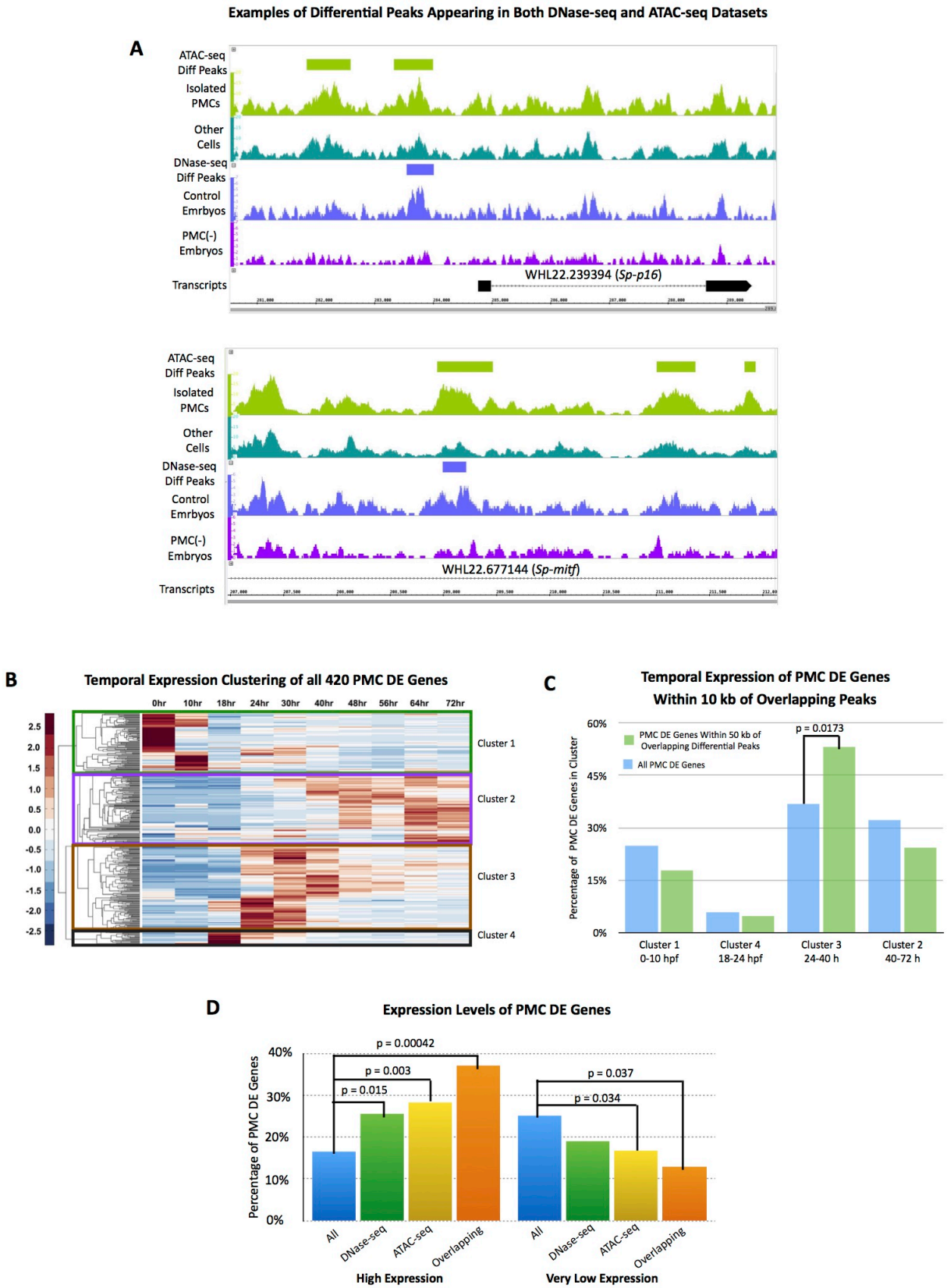


Figure 3.3: Correspondence Between DNase-seq and ATAC-seq Datasets A) ATAC-seq differential peaks (green rectangles) and DNase-seq differential peaks (violet rectangles) located near the *Sp-p16* gene and the *Sp-mitf* gene, both PMC DE genes. Aligned reads, averaged across replicates, from isolated PMCs (light green trace) and other non-PMC cells (dark green trace) using ATAC-seq and control 28 hpf embryos (violet trace) and PMC(-) embryos (dark purple trace) using DNase-seq, are shown. B) Temporal expression profiles (Tu et al., 2012) of 420 PMC DE genes identified previously (Rafiq et al., 2014). Each gene is represented by a single row. The color scale ranges from deep red (2.5-fold higher than mean expression) to deep blue (2.5-fold lower than mean expression). White indicates the mean expression value. Four clusters are delineated, corresponding to maximal gene expression at 0-10, 40-72, 24-40 and 18-24 hpf (hours post fertilization) respectively. C) Temporal expression of the 62 PMC DE genes within 10 kb of overlapping differential peaks: these PMC DE genes were classified into four clusters, delineated in Fig. 3.3B. A significant enrichment was observed in Cluster 3 (24-40 hpf max expression) for PMC DE genes within 10 kb of overlapping differential peaks. D) PMC DE genes were classified into categories based on levels of gene expression in isolated PMCs (data obtained from (Rafiq et al., 2014). "High" expression genes: FPKM between 2512 and 100 (top 17% of all 420 DE genes); "very low" expression genes: FPKM between 14 and 0 (bottom 25% of all 420 DE genes). PMC DE genes with "high" expression levels are significantly enriched in all differential peak sets while PMC DE genes with " very low" expression levels are significantly depleted within 10 kb of the differential ATAC-seq and overlapping peak sets.

We performed ATAC-seq on one batch of 128-cell (11 hpf) *S. purpuratus* embryos to investigate whether putative PMC CRMs were accessible during early cleavage, several hours before the majority of skeletogenic lineage genes are expressed. A large number of overlapping differential peaks (127/161; 79%) were found to be hypersensitive at the 128-cell stage (see Supp. Fig. 3.3 for examples and Supp. Table 5 for the list of overlapping differential peaks). The set of 34 overlapping differential peaks that were not accessible at the 128-cell stage were similar with respect to their position relative to the closest gene, their proximity to DE genes, and the temporal expression profiles of neighboring DE genes, when compared to the set of overlapping differential peaks as a whole.

3.3.4 Differential chromatin accessibility mapping identifies known PMC CRMs

CRMs that regulate four genes expressed selectively by PMCs have been identified by low-throughput approaches and experimentally verified through the mutational analysis of reporter constructs. The four genes are: *Sp-sm50* (Makabe et al., 1995), *Sp-alx1* (Damle and Davidson, 2011), *Sp-tbr* (Wahl et al., 2009) and *Sp-sm30a* (Akasaka et al., 1994; Yamasu

and Wilt, 1999). Each of these CRMs aligned with local regions of open chromatin in one or both of the ATAC-seq and DNase-seq datasets (Fig. 3.4A, B, C, D).

The *Sp-alx1* CRMs (Fig. 3.4A) were not identified as significantly differentially hypersensitive in either dataset. The *Sp-sm50* CRM (Fig. 3.4B) was identified as significantly differentially hypersensitive in the DNase-seq dataset but not in the ATAC-seq dataset, while the *Sp-sm30a* CRM (Fig. 3.4C) was identified as significantly differentially hypersensitive in the ATAC-seq dataset but not in the DNase-seq dataset. Of the four known CRMs involved in regulating *Sp-tbr* expression (Fig. 3.4D), one overlapped an ATAC-seq differential peak and one overlapped both DNase-seq and ATAC-seq differential peaks. These observations showed that known PMC CRMs were well represented in the combined set of differential peaks obtained by DNase-seq and ATAC-seq. They also showed, however, that our identification of PMC CRMs was not exhaustive and that the most complete capture of control CRMs came from combining DNase-seq and ATAC-seq data.

3.3.5 Validation of newly discovered PMC CRMs using GFP reporter gene assays

To validate our experimental and computational identification of PMC CRMs, 31 candidate CRMs were cloned into the EpGFPII plasmid (Cameron et al., 2004) upstream of the *Sp-endo16* promoter (Fig. 3.5A). We focused primarily, but not exclusively, on putative CRMs that were present in both datasets and that were also within 10 kb of PMC DE genes (see Supp. Table 8 for detailed information on all CRMs tested). Reporter plasmids were injected into fertilized *S. purpuratus* eggs and GFP expression was assayed by fluorescence microscopy at 48 hpf. 9/31 constructs (29%) expressed GFP at detectable levels (Fig. 3.5B, Table 3.4). Significantly, all 9 of these reporters drove expression of the reporter gene only in PMCs i.e., none of the constructs we tested resulted in detectable levels of GFP expression in other cell types. The high proportion of active CRMs that showed cell type-specific expression provided a powerful experimental validation of our approach. It should also be noted that the reporter assay was a stringent one which required that a putative CRM was, by itself, sufficient to direct robust, spatially correct expression. Many sea urchin genes are controlled by multiple CRMs, some of which function only to modulate the timing or the level of gene expression (Wahl et al., 2009; Damle and Davidson, 2011; Yuh et al., 1998) , and we would not expect such elements to be active in our assay.

Figure 3.4: Previously Studied PMC-specific Cis-regulatory Modules

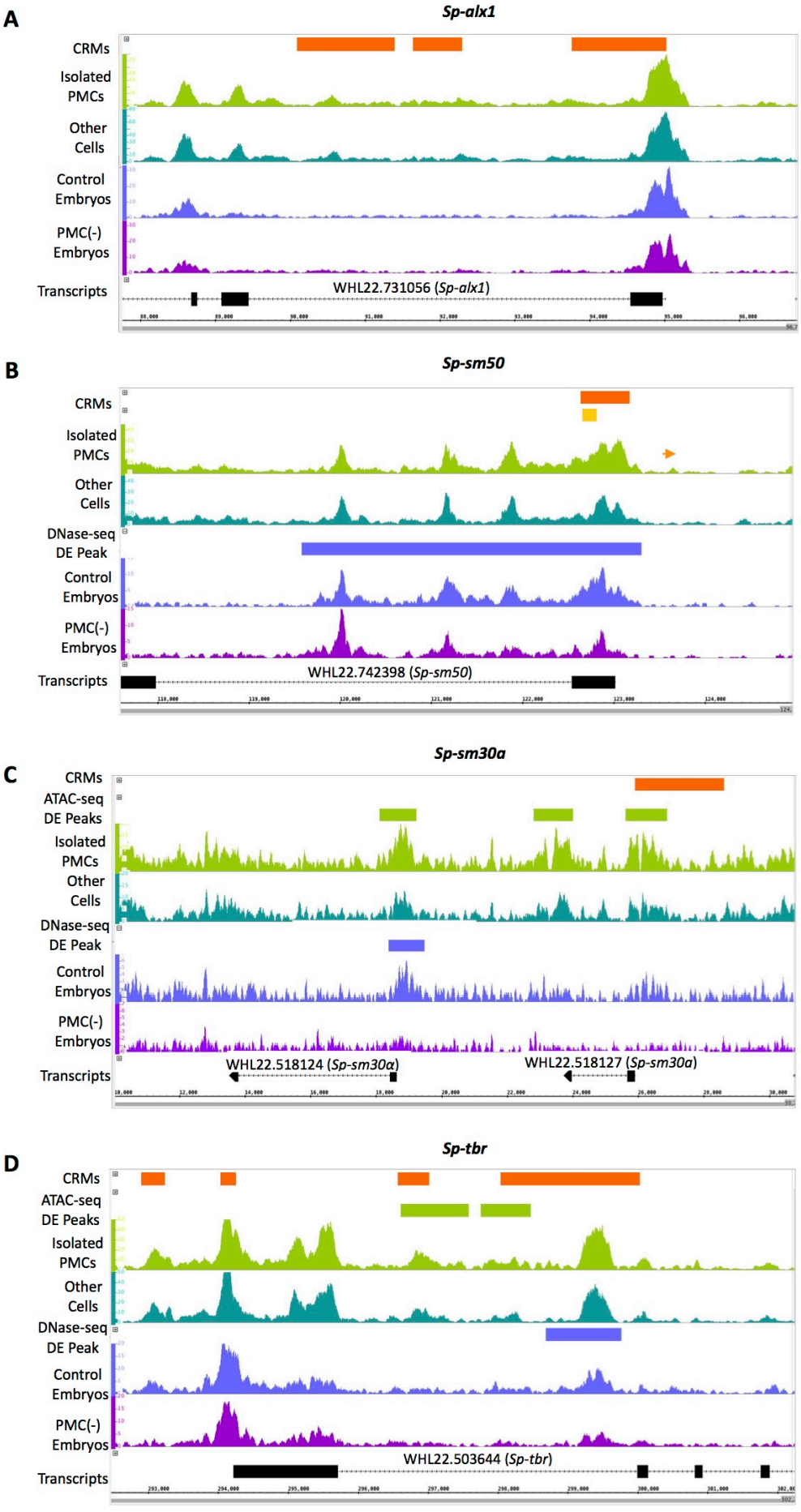


Figure 3.4: Previously Studied PMC-specific Cis-regulatory Modules Previously studied cis-regulatory modules (CRMs) (corresponding to orange and yellow rectangles) critical for the correct spatio-temporal expression of four PMC DE genes are represented in both ATAC-seq (light green trace: isolated PMCs; dark green trace: other cells) and DNase-seq (violet trace: control whole embryos, purple trace: PMC-minus U0126-treated embryos) datasets. A) *Sp-alx1* cis-regulatory modules (Damle and Davidson, 2011) are not represented as differential peaks in the ATAC-seq or DNase-seq datasets. B) The *Sp-sm50* enhancer (orange rectangle) and the minimal element (yellow rectangle) required for correct spatio-temporal expression of *Sp-sm50* (Makabe et al., 1995) are encompassed within a differential peak identified in the DNase-seq dataset (violet rectangle), but not identified as differential in the ATAC-seq dataset. C) The *Sp-sm30a* enhancer (Akasaka et al., 1994; Yamasu and Wilt, 1999) overlaps a differential peak identified in the ATAC-seq dataset (light green), but is not identified as differential in the DNase-seq dataset. D) Two of four previously studied *Sp-tbr cis*-regulatory modules (Wahl et al., 2009) overlap 2 differential peaks in the ATAC-seq (light green) dataset and 1 differential peak in the DNase-seq (violet) dataset.

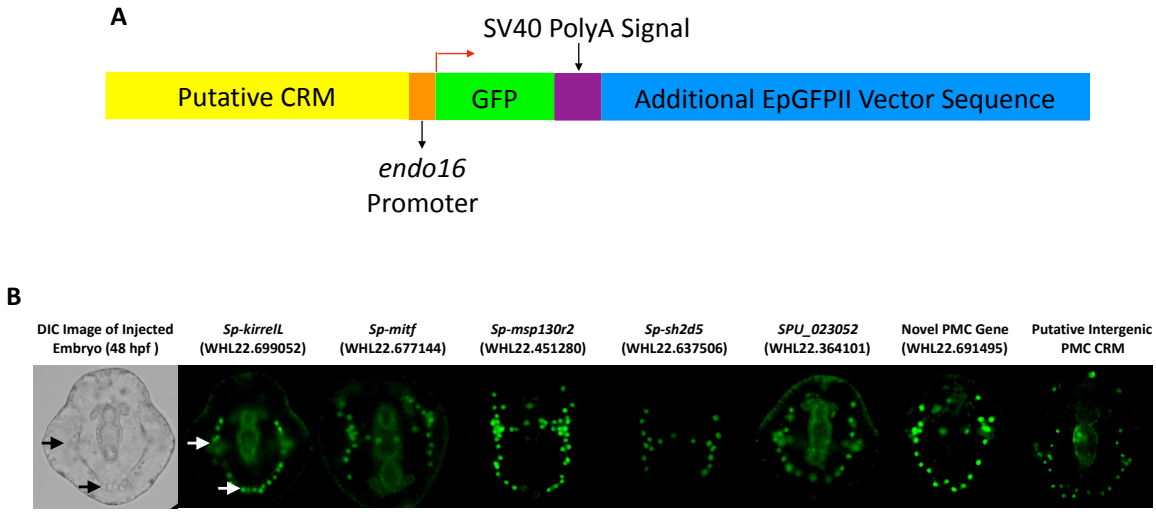


Figure 3.5: Validation of Putative PMC CRMs Using GFP Reporter Gene Assays A) EpGFPII reporter constructs: Of a total of 3,080 PMC-enriched differential peaks identified using DNase-seq and ATAC-seq, 31 peaks were cloned into the EpGFPII plasmid, upstream of the GFP coding sequence and the *Sp-endo16* promoter, and injected into *S. purpuratus* eggs. B) Representative images of *S. purpuratus* embryos injected with 7 reporter constructs, showing PMC-specific GFP expression (green fluorescence) at 48 hpf. Arrows indicate PMCs. DIC and *Sp-kirrelL* images show the same embryo.

Putative CRM Tested	Total Number of Injected Embryos	Number of GFP-Positive Embryos	Number of Embryos Expressing GFP in PMCs	Number of Embryos Expressing GFP Ectopically
Sp-kirrelL (WHL22.699052)	288	88 (31%)	88	0
Sp-mitf (WHL22.677144)	230	170 (74%)	170	0
Sp-msp130r2 (WHL22.451280)	244	106 (43%)	106	0
Sp-sh2d5 (WHL22.637506)	80	43 (54%)	43	0
SPU_023052 (WHL22.364101)	149	28 (19%)	28	0
Novel PMC DE Gene (WHL22.691495)	160	109 (68%)	118	0
Intergenic	78	35 (45%)	35	0
Sp-hypp_2386 (WHL22.239326)	86	53 (62%)	53	0
Sp-c-lectin/PMC1 (WHL22.411805)	257	27 (11%)	27	0

Table 3.4: PMC CRMs validated by reporter gene assays. 9 out of 31 injected reporter constructs showed PMC-specific GFP expression at 48 hpf. No ectopic expression was observed.

3.3.6 Computational prediction of transcription factor binding sites in PMC CRMs

Consensus transcription factor binding sequences have been characterized for fourteen sea urchin transcription factors: Ets1, Alx1, Blimp1, Tbr, Tcf1, Gata, Otx, HesC, bZIP, Sox, Myb, Ot1, Gcm and CBF (see Table 3.5 for consensus sequences and citations).

Transcription Factor	Consensus Sequence	Citation	Enrichment
Ets1	(C/A)GGAA or A(C/A)C(C/A)GGAA(C/G)TA	Damle and Davidson, 2011; Consales and Arnone, 2001	DNase-seq, ATAC-seq and overlapping peaks
Alx1	TAATNNNATTA	Damle and Davidson, 2011	DNase-seq, ATAC-seq and overlapping peaks
Blimp1	G(A/G)AA(C/G)(G/T)GAAA; G(A/G)AA(C/G)AAAN	Yuh et. al., 2004	None
Tbr	AGGTGTGA; AGGTGACA	Jarvel et. al., 2014	None
Tcf1	TTCAAAGG	Yuh et. al., 2004	None
Gata	(C/T)GATA(A/G)	Lowry and Atchley, 2000	None
Otx	TAATC(C/T)	Yuh et. al., 2004	None
HesC	CACGTG or CACGCG	Ochiai et. al., 2008; Smith and Davidson, 2008	DNase-seq peaks
bZIP	GCCGATTCAT	Range et. al., 2007	None
Sox	AACAAT	Range et. al., 2007	None
Myb	YAA(CG/TG)	Range et. al., 2007	None
Ot1	ATGCTAAA	Range et. al., 2007	None
Gcm	ATRCGGGY	Calestani and Rogers, 2010	None
CBF	CCAATT	Dayal et. al., 2004	None

Table 3.5: Enrichment of PMC Transcription Factor Consensus Binding Sites in Differential Peaks. Consensus sequences have been characterized for 14 sea urchin transcription factors. Binding sites for Ets1 and Alx1, two PMC-enriched transcription factors, are signficantly enriched ($p < 0.0134$) in ATAC-seq, DNase-seq and overlapping differential peaks. HesC binding sites are significantly enriched ($p = 0.00156$) in the DNase-seq differential peak set.

AME (McLeay and Bailey, 2010) was used to determine whether any of these sites were enriched in the DNase-seq and ATAC-seq differential peak sets compared to non-differential peaks. First, sites enriched in the differential peaks compared to a shuffled control (the same set of nucleotides in scrambled order) were determined. The same operation was carried out for non-differential peaks, and binding sites that were selectively enriched in only differential peaks were identified. Binding sites for Ets1 and Alx1, two PMC-enriched transcription factors that have direct or indirect inputs into half of the known PMC effector genes in the PMC gene regulatory network (Rafiq et al., 2014), were found to be significantly enriched (p-value <0.0134; Fisher's exact test) in DNase-seq, ATAC-seq and overlapping differential peak sets. Binding sites for HesC were found to be significantly enriched (p-value = 0.00156; Fisher's exact test) in the DNase-seq differential peak set. Blimp1 sites were also enriched in DNase-seq and ATAC-seq differential peaks, but this enrichment was not found to be statistically significant. No enrichment was observed for binding sites of transcription factors that function primarily in other embryonic cell types (e.g., Gcm, Sox and Gata). The enrichment of predicted Ets1 and Alx1 binding sites in all three sets of differential peaks (DNase-seq, ATAC-seq, and overlapping) provided additional support for the validity of our CRM identification.

Validated PMC CRM	Predicted TF Binding Sites
Sp-kirrelL (WHL22.699052)	Sox, Tbr, Gcm, bZIP, Otx, Myb, Ets1
Sp-mitf (WHL22.677144)	Sox
Sp-msp130r2 (WHL22.451280)	HesC, Blimp1, Gata, Sox
Sp-sh2d5 (WHL22.637506)	Ets1, Gata, Gcm
SPU_023052 (WHL22.364101)	None
Novel PMC DE Gene (WHL22.691495)	Myb
Sp-hypp_2386 (WHL22.239326)	Blimp, Alx1
Sp-c-lectin/PMC1 (WHL22.411805)	Otx, Ets1
Intergenic	None

Table 3.6: Predicted TF Binding Sites in Validated PMC CRMs. FIMO (Grant et al., 2011) identified several known sea urchin transcription factor consensus binding sequences in PMC CRMs validated by reporter gene assays.

We used FIMO (Grant et al., 2011) to scan the PMC CRMs validated by reporter gene assays for known sea urchin transcription factor consensus binding sequences. Consensus sequences for Sox, Tbr, Gcm, bZIP, Otx, Myb, Ets1, HesC, Blimp1, Gata and Alx1 were identified (Table 3.6). These are candidate regulators of the validated PMC CRMs that can be tested by targeted mutations. Lastly, we used MEME (Bailey et al., 2009, 2015) for the de novo discovery of motifs enriched in the overlapping differential peak set compared to non-differential peaks. Repeating CT (or GA) motifs were identified to be highly enriched in the overlapping differential peak set. (Supp. Fig. 3.4).

3.4 Discussion

We have shown that differential chromatin accessibility can be used for the efficient, high-throughput identification of CRMs in an early embryonic lineage. We generated a set of high-confidence PMC CRMs by applying two independent approaches to identify regions of chromatin with increased accessibility in PMCs. The first approach used DNase-seq to compare the chromatin accessible profiles of control and PMC(-) (U0126-treated) embryos, while the second used ATAC-seq to compare the chromatin accessibility profiles of purified PMCs and the "non-PMC" cell population. We assembled a robust bioinformatics pipeline that was compatible with both DNase-seq and ATAC-seq data and with the large number of scaffolds in the *S. purpuratus* genome assembly and used this pipeline for the large-scale identification of PMC CRMs.

A variety of evidence supports the conclusion that many of the differential peaks com-

mon to both the ATAC-seq and DNase-seq datasets represent CRMs selectively active in PMCs. First, the peaks in this set were much more likely to lie near genes differentially expressed by PMCs than were other peaks. In addition, genes located within 10 kb of these peaks were more likely than other genes to be associated with biomineralization, the unique developmental function of PMCs. The 62 PMC DE genes near overlapping differential peaks were more likely than other PMC DE genes to be maximally expressed between 24-40 hpf and to be expressed at high levels, both features of the expression patterns of most known biomineralization genes. Indeed, we confirmed that nearly half of the functionally annotated genes in this set have been classified as biomineralization genes. We identified many specific examples of overlapping differential peaks located near well-characterized effectors of skeletal morphogenesis, including *Sp-kirrelL* (Ettensohn and Dey, 2017), *Sp-p16* (Cheers and Ettensohn, 2005), several spicule matrix and MSP130 family genes (Livingston et al., 2006), and carbonic anhydrase (Mitsunaga et al., 1986). Significantly, consensus binding sites for Ets1 and Alx1, key transcription factors that provide regulatory inputs into almost half of all genes differentially expressed by PMCs (Rafiq et al., 2014), were highly enriched in the set of overlapping differential peaks. Lastly, and most importantly, a significant fraction of putative CRMs from this set (6/20; 30%) that we tested experimentally contained sufficient regulatory information to drive reporter gene expression selectively in PMCs, while none supported expression in other cell types. Based on these observations, we conclude that a large proportion of the 161 peaks in the overlapping differential set represent bona fide PMC CRMs.

While differential accessibility is a reliable predictor of PMC CRMs, the converse is not true; i.e., absence of differential signal is not strong evidence that a given region of non-coding DNA lacks regulatory function in PMCs. Of course, some genes are ubiquitously expressed and their CRMs are probably open in all cell types. Even for those genes differentially expressed by PMCs, as discussed below, it seems likely that many of the relevant CRMs are hypersensitive in non-PMC lineages and this may have reduced our ability to detect such regions. We also carried out our analyses at a single developmental stage (28 hpf), and some PMC CRMs might exhibit maximal differential accessibility earlier or later in embryogenesis.

Of the set of previously verified PMC CRMs that we examined (i.e., those that regulate *Sp-alx1*, *Sp-tbr*, *Sp-sm30a*, and *Sp-sm50*) most were identified as differentially open in our analysis. In several cases, however, these CRMs were detected in either the ATAC-seq dataset or the DNase-seq dataset, but not in both. This reinforces the view that a requirement for differential accessibility in both datasets is a very stringent one and points to the reliability of this peak set. At the same time, it indicates that many additional PMC CRMs were identified as differentially accessible by only one of the two approaches. Indeed, 11 CRMs of this type were tested using reporter gene assays and 3 drove GFP expression specifically in PMCs. In addition, when we considered the 3,080 peaks identified as differential by either the ATAC-seq or DNase-seq analysis, nearly 15% of the genes within 10 kb of these peaks were PMC DE genes – a significant enrichment (Fisher's exact test

p-value <2.2e-16; 7.6-fold enrichment). Taken as whole, these considerations suggest that while many bona fide CRMs that drive differential gene expression in PMCs are missing from the highest confidence peak set (i.e., the set of overlapping differential peaks), most are probably contained in one or the other of the two individual datasets.

Although 100% of the PMC CRMs that were active in the EpGFPII reporter plasmid drove expression specifically in PMCs, most of the CRMs we tested did not produce detectable levels of GFP expression. It is important to note, however, that our reporter gene assay required that a cloned CRM function in isolation to direct cell type-specific expression at levels sufficient for detection by fluorescence microscopy. This was a stringent test that could not have detected CRMs that act only to modulate the level or timing of gene expression or those that must cooperate with other regulatory elements to regulate transcription. Thus, a significant fraction of bona fide regulatory elements will lack activity by this assay.

Our findings revealed that, for the most part, the chromatin landscape of PMCs is not highly specific to this cell type. Although we identified reproducible differences in local chromatin accessibility that were predictive of functional CRMs, we rarely observed dramatic differences in peak signals. For example, when we compared the ATAC-seq profiles of purified PMCs and non-PMCs, we consistently observed relatively subtle differences in chromatin accessibility even at CRMs that were subsequently validated experimentally by reporter gene analysis. This strongly suggests that the CRMs of genes expressed specifically by PMCs are open in other cell lineages during early development. It is important to note that we used the same PMC purification method for previous RNA-seq studies and found that FPKM values for PMC-specific mRNAs were typically more than an order of magnitude higher in the PMC fraction than in other cells, confirming the effectiveness of the PMC isolation procedure (Rafiq et al., 2014). To our knowledge, only one other study has compared the accessibility of tissue-specific CRMs in different cell lineages of early embryos. Recently, Pearson et al, (2016) (Pearson et al., 2016) used FAIRE-seq to compare the chromatin accessibility profiles of purified Drosophila embryonic CNS midline cells and intact embryos. Although their study focused on the utility of using differential chromatin accessibility as a enhancer discovery tool, they noted that almost 50% of (9/19) of previously identified enhancers regulating genes expressed preferentially by midline cells had peak signals that were indistinguishable from those in whole embryos. Since midline CNS cells represent a small fraction (<1%) of all embryonic cells, this suggests that these enhancers are open in other cell types.

One explanation for the hypersensitive state of skeletogenic CRMs in non-PMC lineages may lie in the well-known developmental plasticity of sea urchin embryonic cells. Cell types other than PMCs, including endoderm and non-skeletogenic mesoderm cells, have the capacity to adopt a skeletogenic fate under certain experimental conditions, even late in gastrulation (Ettensohn and McClay, 1988; McClay and Logan, 1996; Sharma and Ettensohn, 2011). The skeletogenic potential of these cells may be associated with the priming of PMC CRMs. Surprisingly, our findings strongly suggest that this holds true even of

CRMs that regulate terminal skeletogenic differentiation genes. Studies on pluripotent embryonic stem (ES) cells have identified primed (poised) enhancers that are characterized by open chromatin and other epigenetic marks, yet are transcriptionally inactive (Buecker and Wysocka, 2012). These poised enhancers have been associated primarily with early regulators of cell lineage commitment, and there is very limited evidence that ES cell pluripotency involves protein-DNA interactions at enhancers of terminal differentiation genes (Xu et al., 2009). At present, we cannot determine whether the accessibility of PMC CRMs in other lineages reflects the association of these regulatory elements with transcriptional activators or with repressors. In support of the latter, we detected an enrichment of binding sites for HesC in these CRMs. HesC acts as a repressor of skeletogenic genes and presumably interacts with these sites only in non-PMC lineages, where the protein is expressed (Revilla-i Domingo et al., 2007). We therefore favor the hypothesis that CRMs that regulate terminal skeletogenic effector genes are open in non-PMC lineages as a consequence of their association with HesC and possibly other repressors.

Our ATAC-seq analysis of 128-cell stage embryos showed that most of the high-confidence set of overlapping differential peaks, including several experimentally verified PMCs CRMs, were hypersensitive at the 128-cell (late cleavage) stage, several hours prior to the zygotic activation of skeletogenic effector genes. If enhancer priming reflects a preactivation state, as is widely believed (Zaret and Carroll, 2011; Buecker and Wysocka, 2012), then these findings suggest that pioneer transcription factors interact with PMC CRMs very early in embryogenesis and point to the earliest PMC-specific transcription factors, such as Alx1, as candidates. However, as noted above, hypersensitivity at the 128-cell stage may instead reflect the binding of repressors in non-PMC lineages. Further analysis of purified cell populations will be required to define the temporal and spatial patterns of hypersensitivity exhibited by PMC CRMs during early embryogenesis.

In previous work we identified 420 genes differentially expressed by PMCs, a gene set that included large numbers of terminal effectors as well as several regulatory genes that had not been previously incorporated into the network (Rafiq et al., 2014). We showed that approximately half of the genes differentially expressed in PMCs were regulated by both *Sp-alx1* and *Sp-ets1*, although the mechanism of this co-regulation was not explored. In this study, we found a significant enrichment of both Alx1 and Ets1 binding sites in ATAC-seq differential peaks, DNase-seq differential peaks, and the overlapping peak set, suggesting that a large proportion of PMC CRMs receive direct inputs from both Alx1 and Ets1 (or possibly from other homeodomain and ETS family proteins with very similar binding sites). Because *Sp-alx1* is positively regulated by *Sp-ets1* (Ettensohn et al., 2003; Damle and Davidson, 2011) this suggests that a feedforward mechanism originally proposed by Oliveri and co-workers to account for the regulation of *Sp-msp130*, *Sp-msp103L*, and *Sp-foxb* (Oliveri et al., 2008) may control a large fraction of the effector genes in the PMC GRN. Our studies also point to previously unidentified regulators, as several CRMs active in our reporter gene assay lack consensus binding sites for Alx1, Ets1, or any other transcription factor currently incorporated into the PMC network. In

this regard, we also found using de novo motif searching that poly-CT (poly-GA) tracts are significantly enriched in differential peaks compared to peaks that are not differential. The significance of these low-complexity motifs is unknown, but they may be recognized by sequence-specific DNA-binding proteins such as GAGA-binding proteins, chromatin modifiers that bind preferentially to clustered GAGAG elements and are associated with local nucleosome depletion (Adkins et al., 2006; Berger and Dubreucq, 2012).

Our work has general implications for GRN analysis. The PMC GRN and all other current developmental GRN models have been deduced largely from gene knockdown and gene expression studies and therefore represent networks of functional (epistatic) interactions. Relatively few gene interactions have been characterized at the level of protein-CRM binding and many are undoubtedly indirect. One major barrier to a more complete understanding of GRN architecture has been the challenge of identifying CRMs in a high-throughput manner. In the case of the PMC GRN, for example, only a handful of CRMs were known prior to our work (Makabe et al., 1995; Akasaka et al., 1994; Yamasu and Wilt, 1999; Amore and Davidson, 2006; Wahl et al., 2009; Damle and Davidson, 2011). In the future, it should be possible to combine our strategy with the barcoding of reporter constructs (Nam and Davidson, 2012) to further enhance CRM discovery. Experimental dissection of selected PMC CRMs identified in this study will make it possible to identify key binding sites and direct regulatory inputs.

3.5 Conclusions

We used a combination of DNase-seq and ATAC-seq to identify a high-confidence set of CRMs that regulate gene expression in the skeletogenic lineage. Our work demonstrates the value of using differential chromatin accessibility for the high-throughput identification of CRMs in embryonic tissues. Our identification of hundreds of CRMs selectively active in PMCs has advanced our understanding of the skeletogenic gene network and enhances its value as a model for GRN architecture. This approach for CRM identification can be extended to any embryonic cell type that can be isolated from any organism with a reasonably well assembled genome. Our studies also reveal that several lineage-specific CRMs are hypersensitive hours before most cell type-specific transcripts are expressed, and are open in multiple embryonic cell lineages.

3.6 Methods

S. purpuratus Embryo Culture

Adult *Strongylocentrotus purpuratus* were obtained from Pat Leahy (California Institute of Technology, Pasadena, CA, USA). Gametes were collected from *S. purpuratus* adults by intracoelomic injection of 0.5M KCl and cultured in artificial seawater at 15°C in a 4-liter beaker fitted with a battery-powered stirrer.

DNase-seq Sample Preparation and Sequencing

PMC(-) embryos were produced by treating embryos with U0126, a MEK inhibitor that selectively blocks PMC specification (Fernandez-Serra et al., 2004; Röttinger et al., 2004). Embryos were treated with 10 μM U0126 continuously from the 2-cell stage and sibling control embryos were treated with vehicle (DMSO) alone. For DNase-seq analysis, embryos from three separate matings were collected at 28 hours post-fertilization (hpf); these samples served as biological replicates. Several control and UO126-treated embryos from each batch were immunostained with monoclonal antibody 6a9 (Ettensohn and McClay, 1988) to confirm that PMC specification was effectively blocked (>98%) by U0126 treatment. Nuclei from the three batches of U0126-treated and sibling control embryos were isolated as described by (Coffman and Yuh, 2004).

DNase-seq was performed on isolated nuclei as previously described (John et al., 2013). Briefly, nuclei were digested with 0, 100, 200, 300 and 400 units of DNase I (10 million nuclei per digestion) at 37°C for 3 minutes in digestion buffer (15 mM Tris-HCl pH 8.0, 15 mM NaCl, 60 mM KCl, 0.5 mM EGTA, 1 mM EDTA, 0.5 mM spermidine). The reaction was stopped by adding stop buffer (50 mM Tris-HCl pH 8.0, 100 mM NaCl, 0.1% SDS, 100 mM EDTA, 10μg/ml RNase A, 1 mM spermidine, 0.3 mM spermine) and the digested nuclei were treated with Proteinase K overnight at 55°C. Aliquots of digested nuclei were run on a 0.5% agarose gel, and the digest that produced a light smear (typically a digestion with 200-300 units of DNase I) was selected for further processing.

The selected digests were cleaned by phenol-chloroform extraction, layered on a 9% sucrose solution (0.26 M sucrose, 1 M NaCl, 20 mM Tris-HCl pH 8.0, 5 mM EDTA) and ultracentrifuged in a SW41 swinging bucket rotor at 25,000g for 24 hours at 20°C. 600 μL fractions were collected and 10 μL aliquots were run on a 2% agarose gel, stained with SYBR Green I, and imaged with a Typhoon Gel Imager. Fractions containing DNA fragments <500 bp in size were pooled and mixed with 3X volume of Qiagen QG buffer from the Qiagen MinElute Gel Extraction Kit. 1X volume of isopropanol was added and the samples were purified using Qiagen MinElute columns. Purified DNA was provided to the USC Epigenome Center for library construction (three libraries from PMC-minus

embryos and three from sibling control embryos) and Illumina sequencing (HiSeq2000). Approximately 23.5 million single reads of 50 bp length were obtained per sample.

ATAC-seq Sample Preparation and Sequencing

PMCs and a "non-PMC" cell fraction were isolated from early mesenchyme blastula stage embryos at 24 hpf as described previously (Harkey and Whiteley, 1980; Rafiq et al., 2014). As in this previous study, the purity of the PMC fraction was ∼90% as determined by the fraction of 6a9-positive cells and the depletion of PMCs from the non-PMC fraction was confirmed by RT-PCR. For generating ATAC-seq libraries, PMCs and the corresponding non-PMC fraction were isolated from three embryo cultures derived from separate matings, which served as biological replicates. In one experiment, ATAC-seq was performed on a single batch of 128-cell *S. purpuratus* embryos.

ATAC-seq was performed following the protocol of Buenrostro et al. (Buenrostro et al., 2015) with minor modifications. Briefly, nuclei were extracted from PMCs and other cells by washing three times with lysis buffer (10 mM Tris-HCl, pH 7.4, 10 mM NaCl, 3 mM $MgCl_2$, 0.1% IGEPAL). Nuclei were counted with a hemocytometer. 150,000 nuclei per sample were digested with 2.5 μl transposase (Tn5 transposase from Nextera kit) at 37°C for 30 minutes. The digests were purified using the Qiagen minElute PCR purification kit. The purified DNA was amplified using primers against Illumina adaptors for 5 cycles. The number of additional cycles required for optimal amplification of the library was determined using qPCR. The amplified library was purified using the Qiagen minElute PCR purification kit and provided to the USC Epigenome Center for library construction and sequencing. Six libraries (three PMC libraries and three non-PMC cell libraries) were sequenced with an Illumina NextSeq. Approximately 85 million single reads of 76 bp length were obtained per sample.

Analysis of DNase-seq and ATAC-seq Data

Raw sequence reads were assessed for quality using FastQC (v0.11.4) (Andrews, 2010) and adapter sequences were trimmed using Cutadapt (v1.9) (Martin, 2011). Reads were mapped to the *S. purpuratus* genome using Bowtie2 (v2.1.0)(Langmead and Salzberg, 2012) with default parameters and *S. purpuratus* genome v3.1, obtained from echinobase.org. This is the latest assembly for which a GFF/GTF annotation exists. The v3.1 genome assembly is 826 Mb in size and consists of 32,008 scaffolds with a N50 of 401.6 kb. On average, ∼80% of the reads in each sample were mapped to the genome assembly by Bowtie2.

Samtools (v1.3) (Li et al., 2009) was used to convert the Bowtie2 SAM output format to BAM format. PCR duplications were removed and read counts were equalized using

Samtools. Bedtools (v2.19.1) (Quinlan and Hall, 2010) was then used to convert the BAM output into BED format. The BED files were loaded into Fseq (v1.85) (Boyle et al., 2008) to call peaks using parameters -f 0 and -t 2, where -t 2 is a sensitive peak detection threshold. F-Seq has been shown to be a sensitive and accurate peak caller for DNase-seq and ATAC-seq data (Koohy et al., 2014). The fraction of reads within peaks (the FRiP score) was calculated using Bedtools by extracting and counting all reads within peaks and dividing by the total number of reads mapped. All samples passed a minimum FRiP score threshold of 0.4. Replicate peaks were compared using deepTools (Ramírez et al., 2014) and replicates that were found to be highly concordant (Pearson's correlation coefficient ≥ 0.90) were retained. All DNase-seq replicates met this threshold, but one of three ATAC-seq replicates did not meet the threshold and was not considered for further analysis.

Separate reference peak sets (RPSs) were generated for the DNase-seq and ATAC-seq data by first identifying all replicate peaks that overlapped by at least 75% non-reciprocally and then merging all such peaks across samples separately for the DNase-seq or ATAC-seq data using Bedops (v2.4.2) (Neph et al., 2012). The 75% overlap criterion was enforced non-reciprocally in order to account for differences in peak sizes across replicates. For example, if a 75% or greater overlap was enforced reciprocally, a peak that was >25% larger in one replicate or sample would not have been represented in the RPS. Genome coverage of the reference peak sets was determined by first generating a fasta file containing sequences of peaks in the RPS using Bedtools and then counting the number of nucleotides in the fasta file and dividing this by the number of nucleotides in the *S. purpuratus* genome.

Read counts corresponding to peaks in the RPS were generated using HTSeq (v0.6.0) (Anders et al., 2015) for each replicate. Differential peaks were identified using DESeq2 (Love et al., 2014). Differential peaks in the DNase-seq RPS were identified as peaks that were significantly enriched in the control (whole embryo) replicates compared to the U0126-treated (PMC-deficient) replicates. Peaks were considered significantly enriched if they had nominal p-values <0.1. Differential peaks in the ATAC-seq RPS were identified as peaks that were significantly enriched in the PMC sample compared to the non-PMC sample. Peaks were considered significantly enriched if they had nominal p-values <0.2. A higher p-value threshold was used for ATAC-seq peaks for three reasons: 1) the reduction in the number of replicates (from 3 to 2) compared to the DNase-seq replicates resulted in higher p-values assigned to peaks by DESeq2, 2) one well-characterized PMC CRM in our control set (a CRM that regulates the expression of *Sp-tbr*, see Fig. 3.4) was detected in the differential peak set at a nominal p-value of 0.18 and would have been missed if a lower threshold were chosen and 3) GFP expression in PMCs was observed when the differential peaks around the *Sp-kirrelL* gene (see Fig. 3.2D and Supp. Table 8) with nominal p-values >0.1 were cloned along with the peak with nominal p-value <0.1, but not when this peak was cloned alone. Hence, increasing the p-value threshold to <0.2, we were able to capture additional biologically significant peaks. Nominal, and not adjusted, p-values were used because multiple hypothesis correction was found to be

exceedingly stringent due to the large number of peaks compared.

Overlap between differential peaks identified by DNase-seq and ATAC-seq was determined using Bedops. Differential peaks overlapping non-reciprocally by at least 75% were merged to obtain a set of peaks present in both the ATAC-seq and DNase-seq differential peak sets. Genes within 50 kb of peaks were identified using a custom Python script written by Siddharth Gurdasani. The distribution of peaks with respect to the closest gene and the set of differential peaks within 50 kb of genes differentially expressed by PMCs (DE genes as identified in Rafiq et. al., 2014 (Rafiq et al., 2014) were determined. Peak locations with respect to the nearest gene were defined as follows: Upstream (5′): The 3′ end of the peak was within 1-10 kb upstream of the 5′ end of the first exon; Promoter: The 3′ end of the peak was within 1 kb upstream of the 5′ end of the first exon; Within Gene Body: The 5′ end of the peak was within introns or exons; Downstream (3′): The 5′ end of the peak was within 10 kb downstream of, and did not overlap, the 3′ end of the last exon; Distal: No portion of the peak was within 10 kb of a gene.

128-cell ATAC-seq sequence reads were processed up to the peak-calling stage as described above.

CRM Validation Using GFP Reporter Plasmids

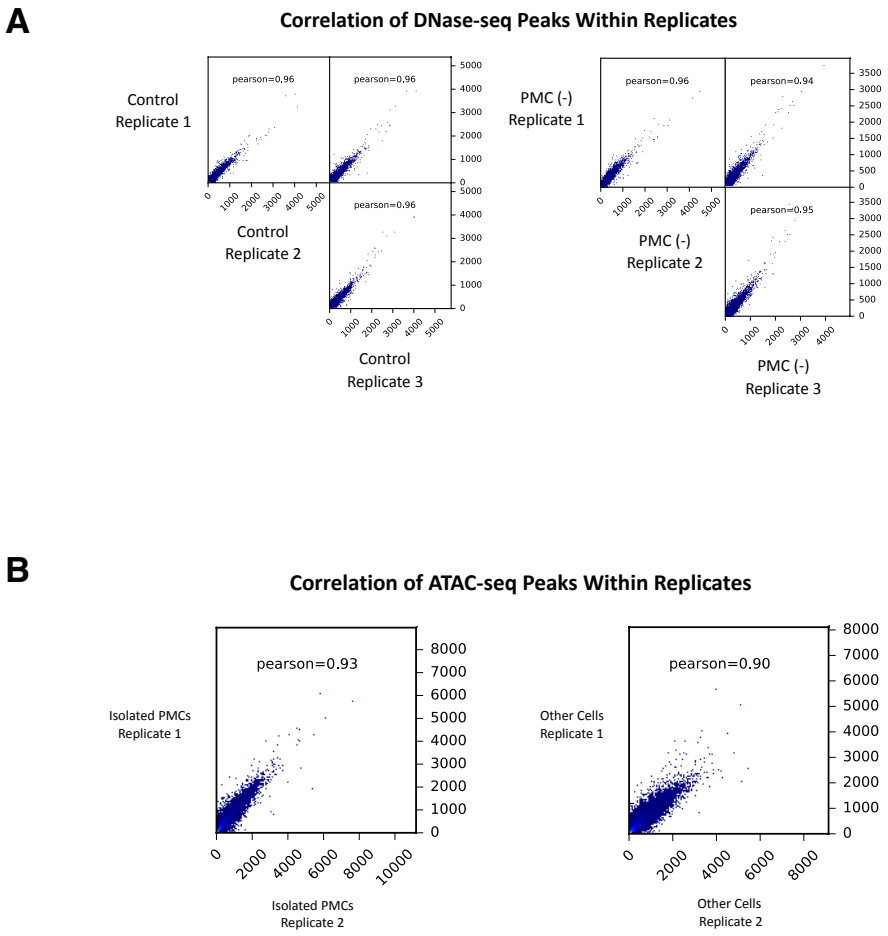
GFP reporter gene constructs were generated by cloning individual, putative PMC CRMs into the EpGFPII plasmid (Cameron et al., 2004). Putative PMC CRMs (see Supp. Table 8) along with ∼200 bp of flanking regions were amplified from *S. purputatus* genomic DNA by PCR and cloned upstream of the basal *Sp-endo16* promoter. In a few cases, adjacent peak regions were also cloned along with the differential peak region. Some constructs also included a promoter peak that was also amplified and cloned upstream of the putative PMC CRM (indicated in Supp. Table 8).

Linearized constructs were injected into *S. purputatus* eggs following established protocols (Arnone et al., 1997, 2004). *S. purpuratus* eggs were fertilized in the presence of 0.1% (wt/vol) para-aminobenzoic acid to prevent hardening of the fertilization envelope. The 20 μl injection solution consisted of 100 ng construct, 500 ng HindIII-digested genomic *S. purputatus* DNA, 0.12 M KCl, 20% glycerol and 0.25% Texas Red dextran. GFP expression was assayed by fluorescence microscopy at the late gastrula stage (48 hpf). Embryos were scored to determine total number of injected embryos (using Texas Red dextran as a marker), the number of embryos showing PMC-specific GFP expression, and the number of embryos with ectopic GFP expression.

Transcription Factor Motif Detection and Analysis

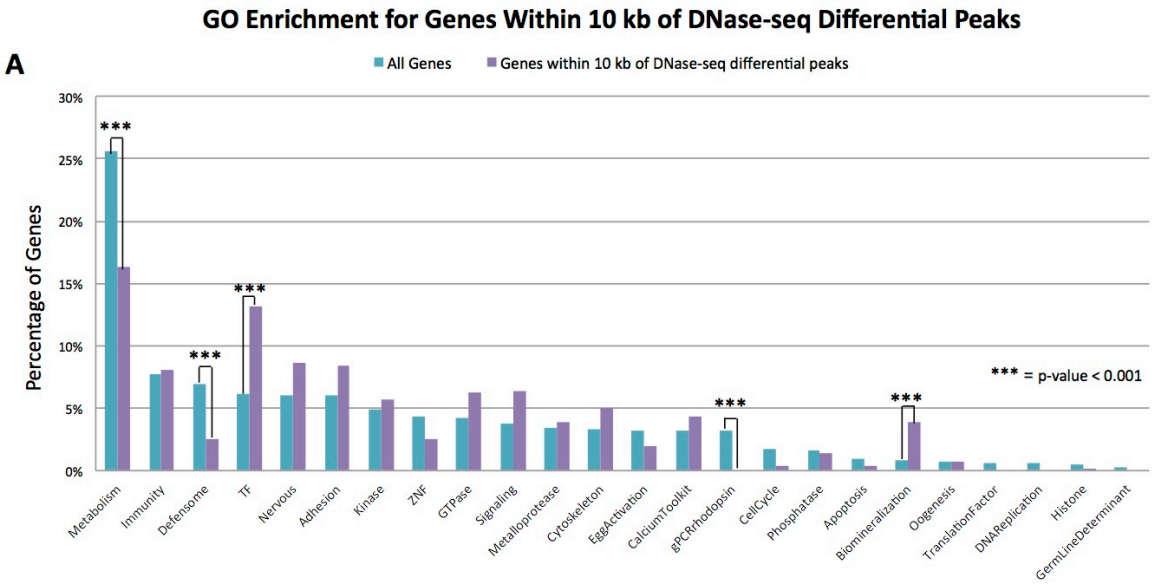
AME (v4.11.2) (McLeay and Bailey, 2010) was used to determine if a set of experimentally verified sea urchin consensus TF binding sites were enriched in differential peaks compared to non-differential peaks. First, enrichment of the consensus TF binding sites in differential peaks compared to a shuffled control was determined. Any sites not also enriched in non-differential peaks compared to a shuffled control were determined to be enriched in differential peaks compared to non-differential peaks. FIMO (v4.11.2) (Grant et al., 2011) was used to search peak sets for sea urchin consensus transcription factor binding sites. MEME (Bailey et al., 2009, 2015) was used for de novo motif searching.

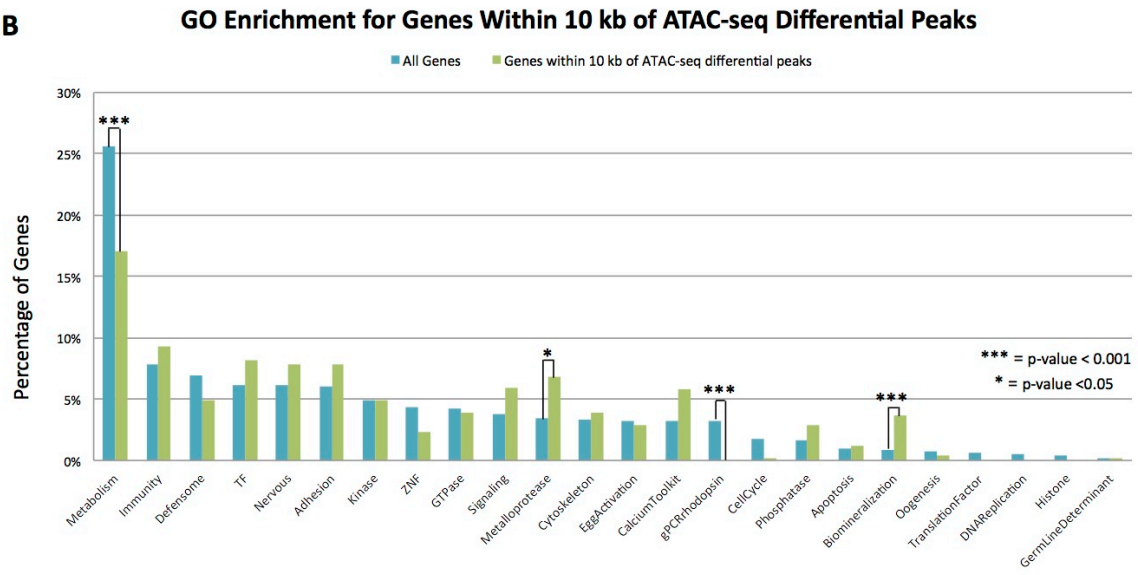
3.7 Supplementary Figures

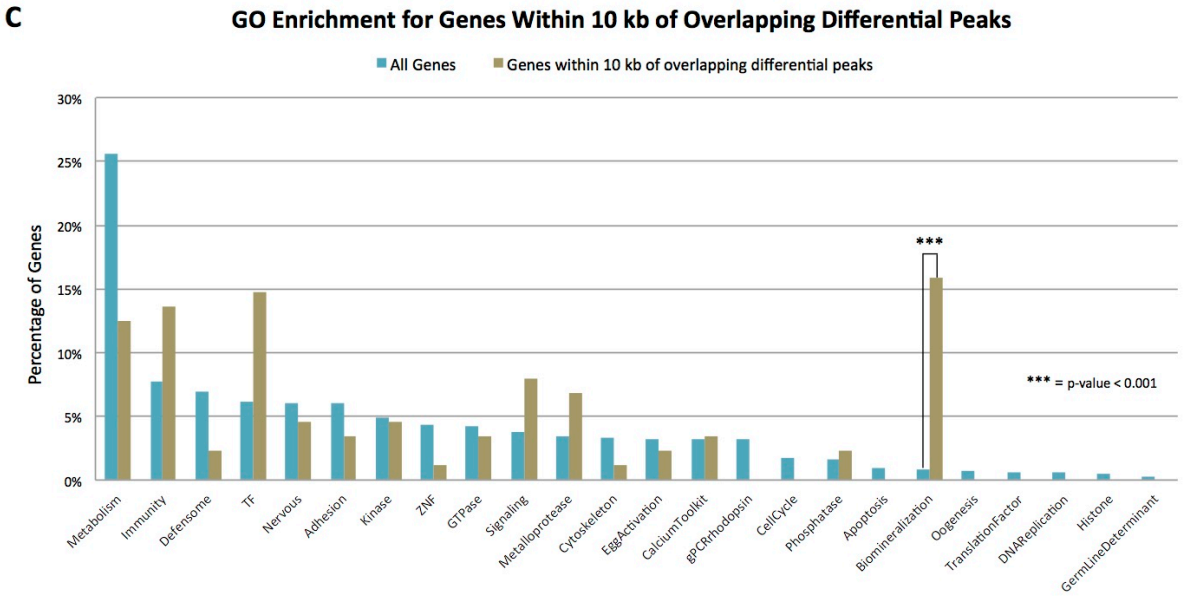


Supplementary Figure 3.1: Correlation of DNase-seq and ATAC-seq peaks within replicates. A) A scatterplot of the read counts of reads aligning to peaks in all three replicates of PMC(-) and control embryos. Replicates are highly concordant, with an average Pearson's correlation of 0.95. B) A scatterplot of the read counts of reads aligning to peaks in replicate 1 and 2 of isolated PMCs and other cells of the embryo. Replicates are highly concordant, with an average Pearson's correlation of 0.915.

Functional Category (GO) Enrichment for Differential Peak Sets

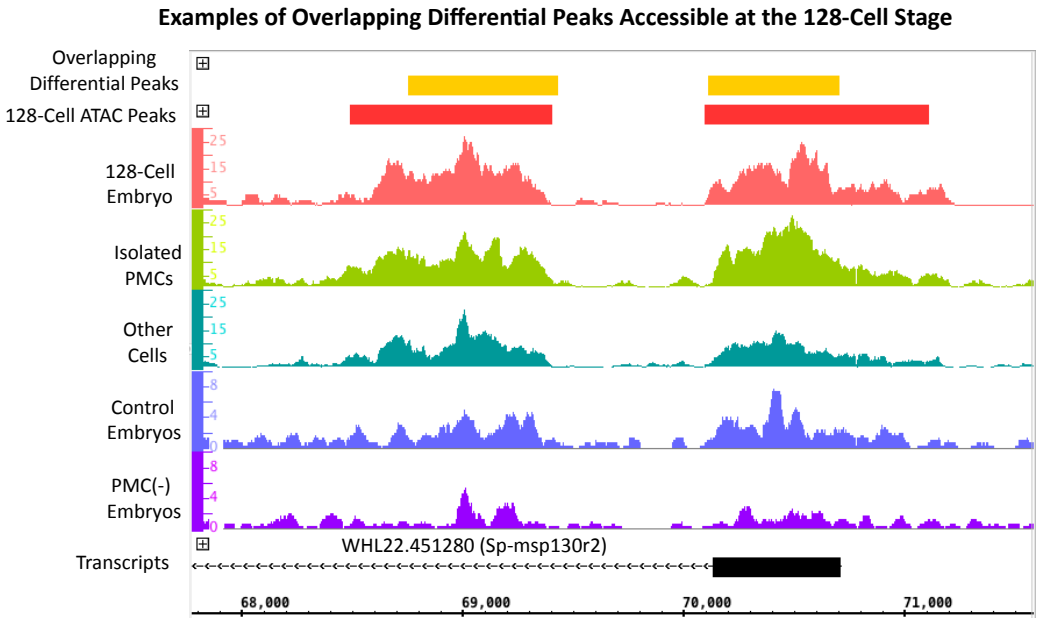






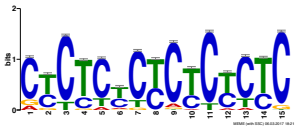
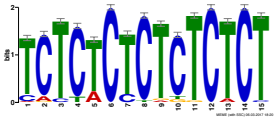
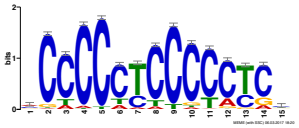
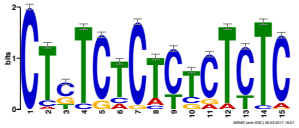
Supplementary Figure 3.2: Functional category (GO) enrichment for differential peak sets. A) The functional categorization of genes within 10 kb of the DNase-seq differential peaks. Functional assignments, obtained from Echinobase are based on hand annotation (Sea Urchin Genome Sequencing Consortium et al., 2006) and on primary GO terms derived by blast2go (Tu et al., 2012). Of the 1,216 genes within 10 kb of differential peaks, 400 have been assigned to functional categories. Genes assigned to multiple functional classes are counted multiple times. B) The functional categorization of genes within 10 kb of the ATAC-seq differential peaks. Of the 1,110 genes within 10 kb of differential peaks, 326 have been assigned to functional categories. C) The functional categorization of genes within 10 kb of the overlapping differential peaks. Of the 135 genes within 10 kb of differential peaks, 55 have been assigned to functional categories.

Examples of overlapping differential peaks accessible at the 128-cell stage.



Supplementary Figure 3.3: Examples of overlapping differential peaks accessible at the 128-cell stage. Overlapping differential peaks (yellow rectangles) around the *Sp-msp130r* gene are accessible at the 128-cell stage (red rectangles represent peaks called at the 128-cell stage). Hypersensitivity corresponding to the overlapping differential peaks is seen at the 128-cell stage (red trace), the 24 hpf stage isolated PMCs (light green trace) and other non-PMC cells (dark green trace), and control 28 hpf embryos (violet trace) and PMC(-) embryos (dark purple trace).

Sequences Enriched in Overlapping Differential Peaks

Motif Logo	No. of Sites	MEME E-value
	119	2.2E-176
	39	2.6E-86
	61	5.9E-66
	40	5.9E-29

Supplementary Figure 3.4: Sequences enriched in overlapping differential peaks, as identified by de novo motif discovery. Four motifs were found to be enriched in overlapping differential peaks compared to non-differential peaks.

Bibliography

Adkins, N.L., Hagerman, T.A., and Georgel, P., 2006. GAGA protein: a multi-faceted transcription factor. *Biochem Cell Biol*, **84**(4):559–67. doi:10.1139/o06-062.

Adomako-Ankomah, A. and Ettensohn, C.A., 2011. P58-A and P58-B: novel proteins that mediate skeletogenesis in the sea urchin embryo. *Developmental biology*, **353**(1):81–93. ISSN 1095-564X. doi:10.1016/j.ydbio.2011.02.021.

Adomako-Ankomah, A. and Ettensohn, C.A., 2013. Growth factor-mediated mesodermal cell guidance and skeletogenesis during sea urchin gastrulation. *Development*, **140**(20):4214–25. doi:10.1242/dev.100479.

Akasaka, K., Frudakis, T.N., Killian, C.E., George, N.C., Yamasu, K., et al., 1994. Genomic organization of a gene encoding the spicule matrix protein SM30 in the sea urchin Strongylocentrotus purpuratus. *J Biol Chem*, **269**(32):20592–8.

Amore, G. and Davidson, E.H., 2006. cis-Regulatory control of cyclophilin, a member of the ETS-DRI skeletogenic gene battery in the sea urchin embryo. *Dev Biol*, **293**(2):555–64. doi:10.1016/j.ydbio.2006.02.024.

Anders, S., Pyl, P.T., and Huber, W., 2015. HTSeq–a Python framework to work with high-throughput sequencing data. *Bioinformatics*, **31**(2):166–9. doi:10.1093/bioinformatics/btu638.

Andrews, S., 2010. FastQC: A quality control tool for high throughput sequence data.

Arnone, M.I., Bogarad, L.D., Collazo, A., Kirchhamer, C.V., Cameron, R.A., et al., 1997. Green Fluorescent Protein in the sea urchin: new experimental approaches to transcriptional regulatory analysis in embryos and larvae. *Development*, **124**(22):4649–59.

Arnone, M.I., Dmochowski, I.J., and Gache, C., 2004. Using reporter genes to study cis-regulatory elements. *Methods in cell biology*, **74**:621–652.

Bailey, T.L., Boden, M., Buske, F.A., Frith, M., Grant, C.E., et al., 2009. MEME SUITE: tools for motif discovery and searching. *Nucleic Acids Res*, **37**(Web Server issue):W202–8. doi:10.1093/nar/gkp335.

Bailey, T.L., Johnson, J., Grant, C.E., and Noble, W.S., 2015. The MEME Suite. *Nucleic Acids Res*, **43**(W1):W39–49. doi:10.1093/nar/gkv416.

Barsi, J.C., Tu, Q., and Davidson, E.H., 2014. General approach for in vivo recovery of cell type-specific effector gene sets. *Genome Res*, **24**(5):860–8. doi:10.1101/gr.167668.113.

Berger, N. and Dubreucq, B., 2012. Evolution goes GAGA: GAGA binding proteins across kingdoms. *Biochim Biophys Acta*, **1819**(8):863–8. doi:10.1016/j.bbagrm.2012.02.022.

Boyle, A.P., Guinney, J., Crawford, G.E., and Furey, T.S., 2008. F-Seq: a feature den-

sity estimator for high-throughput sequence tags. *Bioinformatics*, **24**(21):2537–8. doi: 10.1093/bioinformatics/btn480.

Buecker, C. and Wysocka, J., 2012. Enhancers as information integration hubs in development: lessons from genomics. *Trends Genet*, **28**(6):276–84. doi:10.1016/j.tig.2012.02.008.

Buenrostro, J.D., Wu, B., Chang, H.Y., and Greenleaf, W.J., 2015. ATAC-seq: A Method for Assaying Chromatin Accessibility Genome-Wide. *Curr Protoc Mol Biol*, **109**:21.29.1–9. doi:10.1002/0471142727.mb2129s109.

Cameron, R.A., 2014. Tools for sea urchin genomic analysis. *Methods Mol Biol*, **1128**:295–310. doi:10.1007/978-1-62703-974-1_20.

Cameron, R.A., Oliveri, P., Wyllie, J., and Davidson, E.H., 2004. cis-Regulatory activity of randomly chosen genomic fragments from the sea urchin. *Gene Expr Patterns*, **4**(2):205–13. doi:10.1016/j.modgep.2003.08.007.

Cameron, R.A., Samanta, M., Yuan, A., He, D., and Davidson, E., 2009. SpBase: the sea urchin genome database and web site. *Nucleic Acids Res*, **37**(Database issue):D750–4. doi:10.1093/nar/gkn887.

Cheers, M.S. and Ettensohn, C.A., 2005. P16 is an essential regulator of skeletogenesis in the sea urchin embryo. *Developmental Biology*, **283**:384–396.

Coffman, J.A. and Yuh, C.H., 2004. Identification of sequence-specific DNA binding proteins. *Methods Cell Biol*, **74**:653–75.

Crawford, G.E., Holt, I.E., Whittle, J., Webb, B.D., Tai, D., et al., 2006. Genome-wide mapping of DNase hypersensitive sites using massively parallel signature sequencing (MPSS). *Genome Res*, **16**(1):123–31. doi:10.1101/gr.4074106.

Damle, S. and Davidson, E.H., 2011. Precise cis-regulatory control of spatial and temporal expression of the alx-1 gene in the skeletogenic lineage of s. purpuratus. *Dev Biol*, **357**(2):505–17. doi:10.1016/j.ydbio.2011.06.016.

Duloquin, L., Lhomond, G., and Gache, C., 2007. Localized VEGF signaling from ectoderm to mesenchyme cells controls morphogenesis of the sea urchin embryo skeleton. *Development*, **134**(12):2293–302. doi:10.1242/dev.005108.

Ernst, J., Kheradpour, P., Mikkelsen, T.S., Shoresh, N., Ward, L.D., et al., 2011. Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature*, **473**(7345):43–9. doi:10.1038/nature09906.

Ettensohn, C.A., 2009. Lessons from a gene regulatory network: echinoderm skeletogenesis provides insights into evolution, plasticity and morphogenesis. *Development*, **136**(1):11–21. doi:10.1242/dev.023564.

Ettensohn, C.A., 2013. Encoding anatomy: Developmental gene regulatory networks and morphogenesis. *Wiley Periodicals*, **27**:1–27. ISSN 1526-968X. doi:10.1002/dvg.22380.

Ettensohn, C.A. and Dey, D., 2017. KirrelL, a member of the Ig-domain superfamily of adhesion proteins, is essential for fusion of primary mesenchyme cells in the sea urchin embryo. *Dev Biol*, **421**(2):258–270. doi:10.1016/j.ydbio.2016.11.006.

Ettensohn, C.A., Illies, M.R., Oliveri, P., and De Jong, D.L., 2003. Alx1, a member of the Cart1/Alx3/Alx4 subfamily of Paired-class homeodomain proteins, is an essential component of the gene network controlling skeletogenic fate specification in the sea urchin embryo. *Development*, **130**:2917–2928.

Ettensohn, C.A. and McClay, D.R., 1988. Cell lineage conversion in the sea urchin embryo. *Developmental Biology*, **125**:396–409.

Fernandez-Serra, M., Consales, C., Livigni, A., and Arnone, M.I., 2004. Role of the ERK-mediated signaling pathway in mesenchyme formation and differentiation in the sea urchin embryo. *Dev Biol*, **268**(2):384–402. doi:10.1016/j.ydbio.2003.12.029.

Grant, C.E., Bailey, T.L., and Noble, W.S., 2011. FIMO: scanning for occurrences of a given motif. *Bioinformatics*, **27**(7):1017–8. doi:10.1093/bioinformatics/btr064.

Harkey, M.A. and Whiteley, A.H., 1980. Isolation , Culture , and Differentiation of Echinoid Primary Mesenchyme Cells. *Roux's Archives of Developmental Biology*, **122**(1896):111–122.

Ingersoll, E.P. and Wilt, F.H., 1998. Matrix metalloproteinase inhibitors disrupt spicule formation by primary mesenchyme cells in the sea urchin embryo. *Developmental Biology*, **196**:95–106.

John, S., Sabo, P.J., Canfield, T.K., Lee, K., Vong, S., et al., 2013. Genome-scale mapping of DNase I hypersensitivity. *Curr Protoc Mol Biol*, **Chapter 27**:Unit 21.27. doi:10.1002/0471142727.mb2127s103.

Koohy, H., Down, T.A., Spivakov, M., and Hubbard, T., 2014. A comparison of peak callers used for DNase-Seq data. *PLoS One*, **9**(5):e96303. doi:10.1371/journal.pone.0096303.

Kurokawa, D., Kitajima, T., Mitsunaga-Nakatsubo, K., Amemiya, S., Shimada, H., et al., 1999. HpEts, an ets-related transcription factor implicated in primary mesenchyme cell differentiation in the sea urchin embryo. *Mechanisms of development*, **80**:41–52.

Langmead, B. and Salzberg, S.L., 2012. Fast gapped-read alignment with Bowtie 2. *Nat Methods*, **9**(4):357–9. doi:10.1038/nmeth.1923.

Levine, M. and Davidson, E.H., 2005. Gene regulatory networks for development. *Proc Natl Acad Sci U S A*, **102**(14):4936–42. doi:10.1073/pnas.0408031102.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., et al., 2009. The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, **25**(16):2078–9. doi:10.1093/bioinformatics/btp352.

Livingston, B.T., Killian, C.E., Wilt, F., Cameron, A., Landrum, M.J., et al., 2006. A genome-wide analysis of biomineralization-related proteins in the sea urchin Strongylocentrotus purpuratus. *Developmental Biology*, **300**:335–348.

Logan, C.Y., Miller, J.R., Ferkowicz, M.J., and McClay, D.R., 1999. Nuclear beta-catenin is required to specify vegetal cell fates in the sea urchin embryo. *Development*, **126**(2):345–57.

Love, M.I., Huber, W., and Anders, S., 2014. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol*, **15**(12):550. doi:10.1186/s13059-014-0550-8.

Makabe, K.W., Kirchhamer, C.V., Britten, R.J., and Davidson, E.H., 1995. Cis-regulatory control of the SM50 gene, an early marker of skeletogenic lineage specification in the sea urchin embryo. *Development*, **121**(7):1957–1970.

Martin, M., 2011. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet*, **17**(1):1–10.

McClay, D.R. and Logan, C.Y., 1996. Regulative capacity of the archenteron during gastrulation in the sea urchin. *Development*, **122**(2):607–16.

McLeay, R.C. and Bailey, T.L., 2010. Motif Enrichment Analysis: a unified framework and an evaluation on ChIP data. *BMC Bioinformatics*, **11**:165. doi:10.1186/1471-2105-11-165.

Mitsunaga, K., Fujino, Y., and Yasumasu, I., 1986. Change in the activity of Cl- ,HCO3(-)-ATPase in microsome fraction during early development of the sea urchin, Hemicentrotus pulcherrimus. *Journal of Biochemistry*, **100**:1607–1615.

Nam, J. and Davidson, E.H., 2012. Barcoded DNA-tag reporters for multiplex cis-regulatory analysis. *PLoS One*, **7**(4):e35934. doi:10.1371/journal.pone.0035934.

Natarajan, A., Yardimci, G.G., Sheffield, N.C., Crawford, G.E., and Ohler, U., 2012. Predicting cell-type-specific gene expression from regions of open chromatin. *Genome Res*, **22**(9):1711–22. doi:10.1101/gr.135129.111.

Neph, S., Kuehn, M.S., Reynolds, A.P., Haugen, E., Thurman, R.E., et al., 2012. BEDOPS: high-performance genomic feature operations. *Bioinformatics*, **28**(14):1919–20. doi:10.1093/bioinformatics/bts277.

Oliveri, P., Carrick, D.M., and Davidson, E.H., 2002. A regulatory gene network that directs micromere specification in the sea urchin embryo. *Dev Biol*, **246**(1):209–28. doi:10.1006/dbio.2002.0627.

Oliveri, P., Tu, Q., and Davidson, E.H., 2008. Global regulatory logic for specification of an embryonic cell lineage. *Proceedings of the National Academy of Sciences of the United States of America*, **105**(16):5955–62. ISSN 1091-6490. doi:10.1073/pnas.0711220105.

Pearson, J.C., McKay, D.J., Lieb, J.D., and Crews, S.T., 2016. Chromatin profiling of Drosophila CNS subpopulations identifies active transcriptional enhancers. *Development*, **143**(20):3723–3732. doi:10.1242/dev.136895.

Peled-Kamar, M., Hamilton, P., and Wilt, F.H., 2002. Spicule matrix protein LSM34 is essential for biomineralization of the sea urchin spicule. *Exp Cell Res*, **272**(1):56–61. doi:10.1006/excr.2001.5398.

Peter, I.S. and Davidson, E.H., 2015. *Genomic Control Process: Development and Evolution*. Academic Press.

Peter, I.S. and Davidson, E.H., 2016. Implications of Developmental Gene Regulatory Networks Inside and Outside Developmental Biology. *Curr Top Dev Biol*, **117**:237–51. doi:10.1016/bs.ctdb.2015.12.014.

Quinlan, A.R. and Hall, I.M., 2010. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, **26**(6):841–2. doi:10.1093/bioinformatics/btq033.

Rafiq, K., Cheers, M.S., and Ettensohn, C.A., 2012. The genomic regulatory control of skeletal morphogenesis in the sea urchin. *Development*, **139**:579–590.

Rafiq, K., Shashikant, T., McManus, C.J., and Ettensohn, C.A., 2014. Genome-wide analysis of the skeletogenic gene regulatory network of sea urchins. *Development*, **141**(4):950–61. doi:10.1242/dev.105585.

Ramírez, F., Dündar, F., Diehl, S., Grüning, B.A., and Manke, T., 2014. deepTools: a flexible platform for exploring deep-sequencing data. *Nucleic Acids Res*, **42**(Web Server issue):W187–91. doi:10.1093/nar/gku365.

Revilla-i Domingo, R., Oliveri, P., and Davidson, E.H., 2007. A missing link in the sea urchin embryo gene regulatory network: hesC and the double-negative specification of micromeres. *Proc Natl Acad Sci U S A*, **104**(30):12383–8. doi:10.1073/pnas.0705324104.

Roe, J.L., Park, H.R., Strittmatter, W.J., and Lennarz, W.J., 1989. Inhibitors of metalloendoproteases block spiculogenesis in sea urchin primary mesenchyme cells. *Experimental Cell Research*, **181**:542–550.

Röttinger, E., Besnardeau, L., and Lepage, T., 2004. A Raf/MEK/ERK signaling pathway is required for development of the sea urchin embryo micromere lineage through phosphorylation of the transcription factor Ets. *Development*, **131**(5):1075–87. doi:10.1242/dev.01000.

Sea Urchin Genome Sequencing Consortium, Sodergren, E., Weinstock, G.M., Davidson, E.H., Cameron, R.A., et al., 2006. The genome of the sea urchin Strongylocentrotus purpuratus. *Science*, **314**(5801):941–52. doi:10.1126/science.1133609.

Sharma, T. and Ettensohn, C.A., 2011. Regulative deployment of the skeletogenic gene

regulatory network during sea urchin development. *Development*, **138**(12):2581–90. doi:10.1242/dev.065193.

Smith, J., 2008. A protocol describing the principles of cis-regulatory analysis in the sea urchin. *Nat Protoc*, **3**(4):710–8. doi:10.1038/nprot.2008.39.

Song, L., Zhang, Z., Grasfeder, L.L., Boyle, A.P., Giresi, P.G., et al., 2011. Open chromatin defined by DNaseI and FAIRE identifies regulatory elements that shape cell-type identity. *Genome Res*, **21**(10):1757–67. doi:10.1101/gr.121541.111.

Sun, Z. and Ettensohn, C.A., 2017. TGF-B sensu stricto signaling regulates skeletal morphogenesis in the sea urchin embryo. *Dev Biol*, **421**(2):149–160. doi:10.1016/j.ydbio.2016.12.007.

Thurman, R.E., Rynes, E., Humbert, R., Vierstra, J., Maurano, M.T., et al., 2012. The accessible chromatin landscape of the human genome. *Nature*, **489**(7414):75–82. doi:10.1038/nature11232.

Tu, Q., Cameron, R.A., and Davidson, E.H., 2014. Quantitative developmental transcriptomes of the sea urchin Strongylocentrotus purpuratus. *Dev Biol*, **385**(2):160–7. doi:10.1016/j.ydbio.2013.11.019.

Tu, Q., Cameron, R.A., Worley, K.C., Gibbs, R.a., and Davidson, E.H., 2012. Gene structure in the sea urchin Strongylocentrotus purpuratus based on transcriptome analysis. *Genome research*, **22**(10):2079–87. ISSN 1549-5469. doi:10.1101/gr.139170.112.

Wahl, M.E., Hahn, J., Gora, K., Davidson, E.H., and Oliveri, P., 2009. The cis-regulatory system of the tbrain gene: Alternative use of multiple modules to promote skeletogenic expression in the sea urchin embryo. *Dev Biol*, **335**(2):428–41. doi:10.1016/j.ydbio.2009.08.005.

Weitzel, H.E., Illies, M.R., Byrum, C.A., Xu, R., Wikramanayake, A.H., et al., 2004. Differential stability of beta-catenin along the animal-vegetal axis of the sea urchin embryo mediated by dishevelled. *Development*, **131**(12):2947–56. doi:10.1242/dev.01152.

Wilken, M.S., Brzezinski, J.A., La Torre, A., Siebenthall, K., Thurman, R., et al., 2015. DNase I hypersensitivity analysis of the mouse brain and retina identifies region-specific regulatory elements. *Epigenetics Chromatin*, **8**:8. doi:10.1186/1756-8935-8-8.

Wilt, F.H. and Ettensohn, C.A., 2007. The morphogenesis and biomineralization of the sea urchin larval skeleton. *Handbook of Biomineralization: Biological Aspects and Structure Formation (ed. E. Bauerlein)*, pages 183–210.

Xiong, Q., Zhang, Z., Chang, K.H., Qu, H., Wang, H., et al., 2013. Comprehensive characterization of erythroid-specific enhancers in the genomic regions of human Krüppel-like factors. *BMC Genomics*, **14**:587. doi:10.1186/1471-2164-14-587.

Xu, J., Watts, J.A., Pope, S.D., Gadue, P., Kamps, M., et al., 2009. Transcriptional com-

petence and the active marking of tissue-specific enhancers by defined transcription factors in embryonic and induced pluripotent stem cells. *Genes Dev*, **23**(24):2824–38. doi:10.1101/gad.1861209.

Yamasu, K. and Wilt, F.H., 1999. Functional organization of DNA elements regulating SM30alpha, a spicule matrix gene of sea urchin embryos. *Dev Growth Differ*, **41**(1):81–91.

Yamazaki, A., Kawabata, R., Shiomi, K., Amemiya, S., Sawaguchi, M., et al., 2005. The micro1 gene is necessary and sufficient for micromere differentiation and mid/hindgut-inducing activity in the sea urchin embryo. *Dev Genes Evol*, **215**(9):450–59. doi:10.1007/s00427-005-0006-y.

Yuh, C.H., Bolouri, H., and Davidson, E.H., 1998. Genomic cis-regulatory logic: experimental and computational analysis of a sea urchin gene. *Science*, **279**(5358):1896–902.

Yuh, C.H., Brown, C.T., Livi, C.B., Rowen, L., Clarke, P.J.C., et al., 2002. Patchy interspecific sequence similarities efficiently identify positive cis-regulatory elements in the sea urchin. *Dev Biol*, **246**(1):148–61. doi:10.1006/dbio.2002.0618.

Zaret, K.S. and Carroll, J.S., 2011. Pioneer transcription factors: establishing competence for gene expression. *Genes Dev*, **25**(21):2227–41. doi:10.1101/gad.176826.111.

Chapter 4

Conclusions and Future Directions

My work, as laid out in this thesis, has greatly enhanced the utility of the sea urchin skeletogenic GRN as a model network. I have identified a comprehensive set of regulatory genes and effector genes that carry out skeletal morphogenesis, and we have delineated regulatory inputs into more than half the effector genes identified. In order to establish direct regulatory connections between upstream TFs and downstream skeletal effector genes, I have identified more than 3,000 putative CRMs mediating genes expressed specifically in skeletogenic cells. I have demonstrated the value of using genome-wide techniques to identify relevant genes, locate CRMs mediating the expression of these genes, as well as delineate regulatory inputs into these genes, in a specific embryonic cell lineage. This approach for CRM identification can be extended to any embryonic cell type that can be isolated from any organism with a reasonably well assembled genome.

My thesis work has significantly enhanced our understanding of the PMC GRN, especially with regard to the role and regulation of effector genes, but our understanding of the network is still incomplete. Several open questions have emerged from this work. We have shown than $\sim$50% of effector genes are regulated by Alx1 and/or Ets1, but the upstream regulation of the remaining effector genes is unknown. Four TFs were identified in the RNA-seq study that were not previously known to be enriched in PMCs: Cebpa, Nk7, Mef2 and Mitf. Of these TFs, Mef2 and Mitf are not affected by Alx1 or Ets1 knockdown. In order to determine if these TFs regulate effector genes not regulated by Alx1 and Ets1, morpholinos can be used to knockdown *mef2* and *mitf* expression and RNA-seq can be performed to assay the effect of this knockdown on effector genes. Potential regulatory connections between Mef2 and Mitf and other early and late TFs can also be determined using this method.

While I have been able to identify a large number of putative CRMs regulating PMC effector gene expression, direct regulatory connections between these CRMs and upstream regulators in the network is unknown. Techniques for identifying direct regulatory inputs into CRMs have so far relied on detailed mutational analysis using reporter gene

assays. This approach is time-consuming and laborious. While attempts have been made to increase the throughput of these assays, several bottlenecks still remain, restricting the scope of these methods. To establish direct TF-effector gene linkages, TF binding sites must be identified within effector gene CRMs. A few different high-thoughput methods can be used to identify TF binding sites. Genome-wide footprinting using DNase 1 can identify local regions of DNA protected from cleavage that represent functional TF binding sites. Protein binding microarrays can be used to identify TF binding sites of purified TFs in vitro, and ChIP-seq can also be used to identify potential TF binding sites provided an antibody for the TF of interest is available. Computational approaches can be used to find enriched motifs within CRMs, and exisiting TF motif databases can be referenced to identify the associated TF. Once TF binding sites have been identified in a high-thoughput manner, mutational analysis of these sites can be streamlined by using barcoded GFP assays to validate binding sites identified.

The chromatin profiles I obtained from whole sea urchin embryos as well as isolated PMCs and other cells of the embryo have opened up several interesting avenues for further study. Whole-embryo ATAC-seq data for several developmental stages starting from the blastula stage is available on Echinobase and ATAC-seq can also be performed on early stage embryos. Taken together, chromatin profiles of sea urchin embryos across development can be used to identify putative CRMs regulating key developmental genes, and these CRMs can be followed across development and correlated with gene expression. We can determine when these CRMs first emerge, and how long they remain active. Preliminary ATAC-seq data on an early cleavage stage embryo revealed that several PMC CRMs are accessible even at this early stage, several hours before most skeletogenic genes are expressed. We can focus on CRMs in the PMC lineage by isolating micromeres and culturing them in vitro. Micromeres can autonomously be specified to form PMCs, and with the addition of horse serum, PMCs are able to create skeletal spicules in vivo. If ATAC-seq is performed on isolated micromeres as well as cultured PMCs at various stages, we can follow PMC lineage CRMs across development.

My study on differential chromatin accessibility in isolated PMCs compared with other cells revealed some surprising insights on the plasticity of chromatin during early development. I found that there was very little difference in chromatin accessibility in isolated PMCs compared with other cells. In fact, even CRMs mediating the expression of terminal effector genes were accessible in the other cells to the same degree as in PMCs. It would be interesting to study this further to see why this is the case. It is possible that these CRMs are accessible in other cell lineages because they are bound by repressors. We know, for example, about HesC repressing skeletogenic gene expression in other cells of the embryo. It is also possible that many CRMs are poised for activation at this early developmental stage. We know that the PMC GRN can be activated in a few other cell types of the embryo, so it is likely that CRMs mediating skeletogenic gene expression remain accessible in these other cells. If ATAC-seq is performed on other isolated cell types such as ectodermal cells, gut cells and non-skeletogenic mesoderm cells, we can see the extent

to which these CRMs are active in multiple cell types across development. It is possible to isolate these other cell types of the embryo using FACS sorting as demonstrated by Barsi et. al., 2014.

Our current view of the skeletogenic GRN is relatively static since perturbation analyses have focussed on knocking down genes right from the beginning of development. However, some developmental genes may have different roles at different developmental stages. For example, Alx1 has a role in the early specification of the skeletogenic lineage as well as in biomineralization in the juvenile sea urchin. In order to dissect changes in gene regulation as development proceeds, it is important to perturb gene function at specific times during development. This can be achieved using photoactivable morpholinos or inducible CRISPR-cas9 systems that enable the perturbation of gene function at defined times.

By using the approaches described here, we can significantly improve the comprehensiveness of the skeletogenic gene network in the near future and enhance its status as a model GRN that can be used to answer fundamental developmental and evolutionary questions.