A study of information processing in the sea urchin embryo by rewiring mesodermal gene regulatory networks and cis-regulatory analysis of skeletogenic regulators

Thesis by

Sagar S. Damle

In Partial Fulfillment of the Requirements

For the Degree of

Doctor of Philosophy in

Biology

California Institute of Technology

Pasadena, CA

2011

(Defended January 27, 2011)

ACKNOWLEDGEMENTS

I would like to thank a number of people for helping to make this thesis possible. First I am grateful to my advisor, Eric Davidson, for his mentorship and support. He has been both my harshest critic and my most supportive teacher, and in many ways I'm far more thankful for the former than the latter. I would also like to thank the members of my thesis committee, Dr. Barbara Wold, Dr. Ellen Rothenberg and Dr. Paul Sternberg for their guidance and for their incredibly valuable career advice.

I also want to thank past and present colleagues from the Davidson Lab: my first lab mentor, Takuya Minokawa, for his patience and guidance and for teaching me how to do whole-mount in-situs; and to the post-docs, Cathy Yuh, Paola Oliveri, Andy Ransick, Jongmin Nam, Joel Smith, Qiang Tu, Smadar Ben Tabou De Leon, and fellow grad-students Roger Revilla, Pei Yun Lee, Titus Brown, and Stefan Materna for their support, both as biologists and friends, as well as for many inspirational discussions. Special thanks go to Jane Rigg and Deanna Thomas for providing such a wonderful environment in which to learn. I also want to thank members of the Wold Lab, Chris Hart, Gilberto Hernandez, and Brian Williams for their valuable advice, guidance and friendship.

Finally, I would like to thank my friends and family. Sandy Sharp for reminding me why I love science. Nahoko Iwata, for reminding me what “hard work” *really* means. To my parents, Subhash and Rajashri Damle, to my brother and sister, Shirish and Reshama, and to ‘Buggy’: I am indebted to you all for your support and acceptance over the course of this long journey.

ABSTRACT

The gene regulatory networks (GRNs) specifying embryonic skeletogenesis and pigment cell differentiation in the purple sea urchin, *Strongylocentrotus purpuratus* make predictions about the necessity of regulatory gene expression and cis-regulatory wiring for directing development. Here, these predictions are tested in a novel way, by adding new regulatory linkages to the GRN, effectively rewiring development at the level of genomic DNA. The outcome of this perturbation confirmed the sufficiency of the regulatory factor, *gcm,* to direct pigment cell differentiation, but also identified previously unknown repression functions for *gcm* on *alx1,* an important regulator of skeletogenesis. These results motivated a complete cis-regulatory analysis of *alx1* that identified a potential mechanism for *gcm* repression. Finally, this work describes a method for measuring GFP reporter activity in live sea-urchin embryos that will permit real-time cis-regulatory analysis.

TABLE OF CONTENTS

INTRODUCTION

Specification is the process whereby cells attain different developmental states through a succession of regulatory events that are mediated by transcription factor expression and signaling systems. In all bilaterians specification pathways transform the fertilized egg from a single cell to a multicellular, triploblastic organism composed of several tissues with distinct functions. It is now understood that the instruction sets for specification are essentially hardcoded into genomic DNA in two forms: i) in the coding sequence of regulatory genes which is transcribed and translated by the cell to produce regulatory proteins, whose functions are ultimately to lay down the body plan of the organism and ii) in noncoding regions, as cis-regulatory DNA which integrates regulatory inputs in a given cell and computes when and where genes will be expressed (Davidson 2006). Genomic DNA therefore directs development by deciding, within each and every cell and at every point in time, which proteins should be made and how those proteins will control future developmental events. Furthermore modern molecular biology has shown that specification is a complex event, caused by successive rounds of refinement that are mediated by a network of interactions by regulatory genes acting at multiple points in development. These observations have led to the construction of gene regulatory networks (GRNs) whose structure, we are now seeing, provides system level explanations for developmental and physiological functions (Davidson, 2010).

The sea urchin has been appreciated as a model for how the genome controls development for over a century, but only in the past 20 years have the tools existed to understand development at the molecular level. Currently the GRN describing endomesoderm specification from egg to early gastrula is well understood (Davidson et al., 2002; Peter and Davidson, 2009) and networks for ectoderm specification and for the later development of gut and mesoderm cell types are being rapidly solved (E. Li, S. Materna and I. Peter unpublished). The purpose of this introduction is to summarize key features of mesoderm specification in the sea urchin embryo from the point of view of its underlying gene regulatory network. The sea urchin (Davidson et al., 1998) embryonic mesoderm is composed of a multitude of cell types which include: a skeletogenic mesenchyme, pigment cells, muscle cells, blastocoelar cells, and cells which make up a subset of lateral coelomic pouches and which contribute to mesoderm of the adult body plan.

The sea urchin as a model system for understanding gene regulatory networks

The sea urchin embryo is a useful model system for studying molecular events underlying embryonic bilaterian development for several reasons: the embryo is transparent and has a relatively simple body plan and is only 1 cell-layer thick and comprised of less than 15-20 cell-types; specification occurs before cell migration takes place, thus reducing the complexity of cell-cell signaling interactions to mainly short-range or cell-autonomous cues; it is relatively simple to simulate genetic manipulation (knockouts/overexpression) within the embryo by injection of morpholino antisense RNA or mRNA message, or to perform cis-regulatory analysis using DNA constructs containing reporters fused to cis-regulatory elements; eggs are fertilized externally and can be cultured in large numbers; it is fairly easy to isolate sufficient quantities of protein and mRNA for DNA-binding studies and for measuring spatial and temporal gene expression patterns.

The sea urchin genome was sequenced in 2006 and has offered a wealth of information regarding sea urchin biology. The genome is roughly 800 megabases and encodes on the order of 23000 genes. Analysis of gene categories has revealed the sea urchin contains representatives of nearly all bilaterian transcription factor families and a majority of signal transduction genes (Sodergren et al., 2006). Detailed temporal and spatial embryonic expression profiles for a majority of transcription factors, and zinc finger genes have now been measured (Howard-Ashby et al., 2006a; Howard-Ashby et al., 2006b; Materna et al., 2006) making it possible to estimate the composition of regulatory inputs expressed in nearly all cell types and developmental states in the early embryo. These advances have greatly assisted the pace in which gene regulatory networks for development can be assembled, as evidenced by the rapid pace in which the ectodermal GRN is now being described (Su et al., 2009).

An overview of embryogenesis: from egg to gastrula stage

Embryogenesis, from fertilization to early gastrulation, in *S. purpuratus* embryos has been reviewed extensively (Davidson et al., 1998). The first two cleavages produce 4 macromeres which are essentially developmentally equivalent and each capable of producing a pluteus larva with an oral aboral axis. The third cleavage, however, is orthogonal to the first two and along what is known as the animal/vegetal (A/V) axis and it generates 8 cells of equal size but of differing developmental potential. The embryo at this point can be divided into two tiers of daughter cells. The tier of cells in the animal half of the embryo will ultimately form the ectodermal tissues of the embryo and the tier in the vegetal half will form endodermal and mesodermal tissues. The fourth cleavage also occurs along the A/V axis however here the vegetal-most macromeres divide unequally, producing 4 large micromeres that reside at the vegetal pole and 4 macromeres. With respect to the micromeres, the 5th cleavage is also an unequal division. This cleavage produces 4 small micromeres that remain at the vegetal pole and 4 large micromeres that lie directly above. By the 6th cleavage, or roughly 7-8 hours post fertilization (hpf), vegetal hemisphere macromeres will have divided both laterally and equatorially into two tiers of roughly 16 cells each called veg1 and veg2. The veg2 tier lies adjacent to the micromeres whereas the veg1 tier lies closer to the animal half of the embryo. Detailed lineage tracing experiments have shown that by the end of 6th cleavage, or the 60-cell stage, a number of cell types in the embryo have already undergone specification. These include the cells of the animal hemisphere which will become either oral and aboral ectoderm, the cells of the veg1 and veg2 tier which become endoderm and endomesoderm respectively, the 8 large micromeres which become the skeletogenic mesenchyme (SM), and the small micromeres which do not participate in embryogenesis and which are thought to act as germ cells. From 7th to 9th cleavage a signal from the large micromeres induces specification of non-skeletogenic mesoderm (NSM) in the adjacent veg2 tier. At the same time veg2 cells elongate along the A/V axis and organize at the vegetal pole into a thick, flat structure called the vegetal plate. The center of the vegetal plate contains presumptive mesoderm cells whereas the periphery contains presumptive endoderm. At roughly 20 hpf, or late blastula stage, the large micromeres ingress into the blastocoel and start to differentiate into skeletal cells. As they are the first to ingress, the cells of the SM lineage are also known as primary mesenchyme cells (PMCs). After entering the blastocoel PMCs segregate into two ventrolateral clusters near the base of the archenteron and begin to lay down the embryonic skeleton. Post-ingression skeletogenic cells also fuse to form a syncytium, allowing an exchange of cytoplasmic material and enabling relatively nonmosaic expression of all skeletal genes with the exception of some genes that are specifically localized at the sites of spiculogenesis. The remaining cells of the vegetal plate, the veg2 tier of cells and their veg1 neighbors invaginate at 30 hpf to form the archenteron. Differentiating NSM delaminate from the tip of the archenteron from 30 hpf onward. By late larval stage both SM and NSM cell types have completed their respective differentiation programs to produce an "easel-shaped" skeleton and a wide variety of mesodermal cell types respectively.

Initial anisotropies and maternal inputs effecting mesoderm specification

The process of mesoderm specification involves a series of cell-state bifurcations, the earliest of which separates the SM and NSM lineages. The cascade of events that lead to this initial split is triggered by maternal anisotropies that become compartmentalized during cleavage. By 4th cleavage there is a vegetal bias towards nuclear beta-catenin, a target of the wnt signaling pathway, with the highest nuclear levels in the micromeres of the vegetal pole. This bias is caused by localization of the docking protein Disheveled (Dsh) to the vegetal cortex of the oocyte (Kumburegama and Wikramanayake, 2008; Leonard and Ettensohn, 2007) (Chuang et al., 1996). Dsh acts to block the degradation of cytoplasmic beta-catenin. Cytoplasmic beta-catenin will transit to the nucleus and complex with the transcription factor Tcf to convert it from a repressor to an activator. At the same stage, soxB1 becomes nuclearized in all blastomeres except the micromeres. This is significant because soxB1 is thought to antagonize beta-catenin nuclearization, further sharpening the boundary of beta-catenin localization along the micromere/macromere border (Kenny et al., 1999). In addition to nuclear beta-catenin, otx is also nuclearized with a bias towards the vegetal pole (Chuang et al., 1996) (Smith and Davidson, 2009). It was shown that the combination of beta-catenin nuclearization and maternal otx directly activate expression of pmar1 in the large micromeres (Smith and Davidson, 2009), which is a key event in the specification of the SM lineage.

Pmar1 is a member of the Paired family of homeodomain transcription factors and acts as a repressor. Its expression permits the expression of the set of transcription factors and signaling components required to initialize the skeletogenic regulatory state and to direct the specification of NSM in the adjacent veg2 tier of cells. *Pmar1* acts to repress the expression of a second globally expressed repressor, *hesc*, in what has been described as a double-negative gate(Oliveri et al., 2002; Revilla-i-Domingo et al., 2007). Hesc in turn represses the transcription of *alx1, ets1* and *tbrain* (Revilla-i-Domingo et al., 2007). As a consequence of pmar1 expression, these three transcription factors are expressed in the large micromeres and together make up an initial regulatory state that is necessary for skeletogenesis. Hesc is also a direct target of the Delta ligand. Delta/Notch signaling is required in Veg2 cells to initiate NSM-specific regulatory gene expression, as we will see later.

Pmar1 and *hesc* form an interesting pair of genes and reveal the first of many aspects about network organization that deepen our understanding of developmental processes. The double negative gate serves two purposes in the embryo, the first of which is to allow skeletogenesis to occur in the large micromeres, and the second of which is to prevent this specification state at all other cells. One could imagine a simpler method for specifying the skeletogenic state, i.e. the micromere-specific expression of an activator that induces expression of early skeletogenic initiators, however this mechanism would not have the added advantage of locking out this pathway in other cells. The structure of the double negative gate, and the near-ubiquitous expression pattern of the repressor *hesc* imply that positive regulatory inputs for driving skeletogenesis must be expressed globally in the embryo, either independently of *hesc* expression or downstream of *hesc*. This idea was confirmed when the injection of mRNA encoding *pmar1* or the knockdown of *hesc* was sufficient to convert all blastomeres to a skeletogenic mesenchymal cell state (Revilla-i-Domingo et al., 2007). Recent work has identified the inputs for the *tbrain* (Wahl et al., 2009) and *delta* genes (Smith and Davidson, 2008) and the activator of *alx1* is identified in Chapter 2.

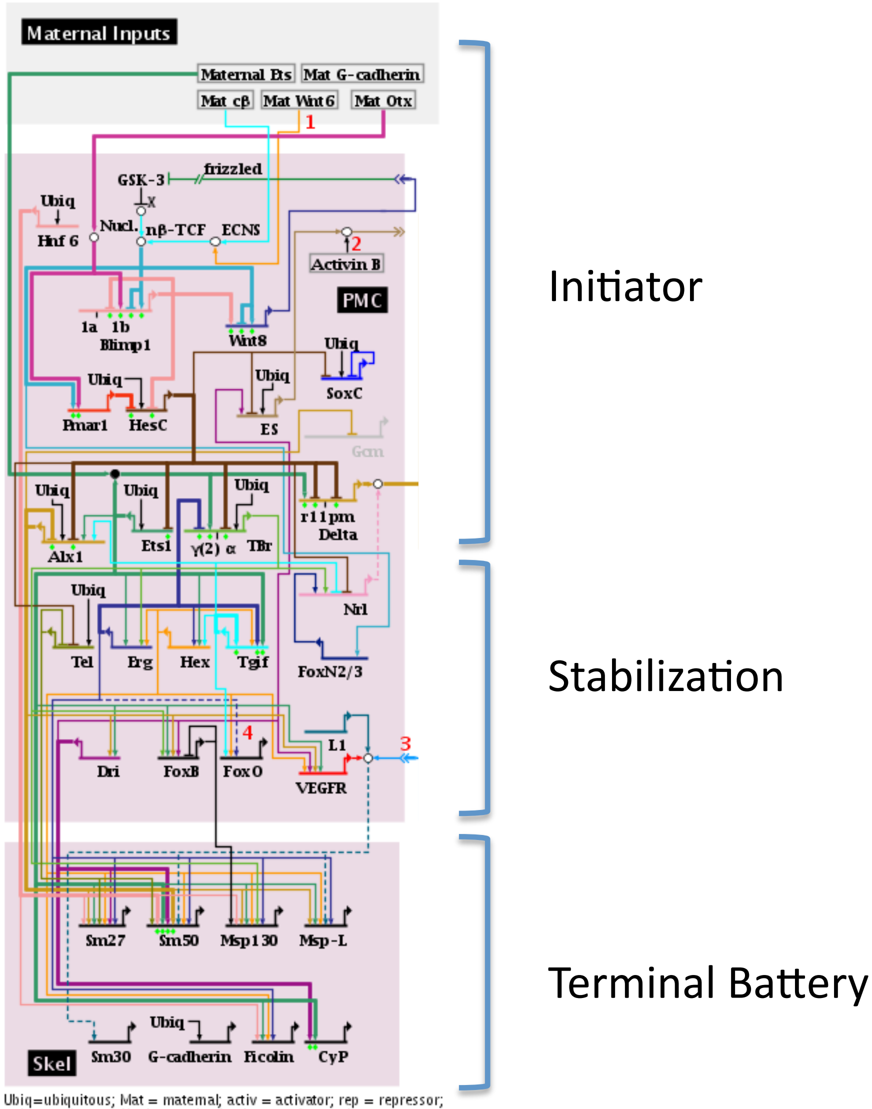
THE SKELETOGENIC REGULATORY STATE

Experiments by Okazaki (Okazaki, 1975) showed that the SM lineage is capable of autonomously running the developmental program required to produce skeletal rods. A dissection of the skeletogenic regulatory network reveals that, beyond the double-negative gate, there are three subcircuits connected in tandem: i) an initiator circuit consisting of genes directly downstream of the double-negative gate, ii) a stabilization circuit that uses extensive autoactivating feedback wiring and iii) a circuit driving terminal differentiation genes. I describe these circuits below in greater detail and along with the relevant regulatory factors.

Initiators of skeletogenesis

After establishment of the double-negative gate *alx1, ets1, tbrain and tel* are the earliest transcription factors to be expressed in the large micromeres. The SM network shows how these genes perform both discreet and collaborative roles in specification (Figure 0.1). Following is a brief summary of the expression pattern and functions of these regulators.

Alx1 was identified as the first invertebrate member of the Cart1/Alx3/Alx4 family of Paired-class homeodomain proteins (Ettensohn et al., 2003). In vertebrates, *cart1, alx3, alx4* and *prx* compose a subgroup of Paired class genes called ‘Group-I aristaless-like genes’, that are expressed during embryogenesis in mesenchyme of craniofacial primordial and limb buds and loss of function mutants of these genes in mice cause developmental defects in skeleton and their embryonic precursors.



Ubiq=ubiquitous; Mat = maternal; activ = activator; rep = repressor;
unkn = unknown; Nucl. = nuclearization; χ = β-catenin source;
nβ-TCF = nuclearized b-β-catenin-Tcf1; ES = early signal;
ECNS = early cytoplasmic nuclearization system; Zyg. N. = zygotic Notch

Figure 0.1. Skeletogenic specification and terminal differentiation gene regulatory network. Maternal anisotropies set up the initiation of a double negative gate that leads to the clearance of the repressor *hesc* and expression of skeletogenic regulators *alx1, tbrain* and *ets1/2* as well as the signaling ligand *delta* which is required for specification of nonskeletogenic mesenchyme. The network can be divided into three subcircuit, an initiator circuit, a specification state stabilizer with dynamic autoregulatory wiring and a terminal differentiation battery.

Alx1 is first expressed after 5th cleavage exclusively in the large micromeres and is essential for driving the regulatory subcircuit that mediates epithelial-to-mesenchymal transition of the SM lineage at 22-24 hpf as well as for driving late regulators of terminal differentiation (*foxB* and *deadringer*) as well as the terminal differentiation battery itself. Though its spatial pattern of expression is straightforward, the expression kinetics of *alx1* are surprisingly complex. Expression peaks rapidly at 10-12 hours post fertilization (hpf) and drops roughly 3 fold by 16hpf, during early blastula stage, and peaks again at roughly 24hpf, during the mesenchyme blastula stage. The timing of the first peak of expression is characteristic of SM-specific genes regulated by the double-negative gate. Indeed SM-specific the expression of *tbrain* and the signaling gene *delta* has been well studied (Smith and Davidson, 2008; Wahl et al., 2009) – *Hesc* has been shown to directly repress these genes throughout the early embryo, and their positive drivers have been identified (ets1 and runx1 respectively) – however little has been demonstrated in regards to the interactions that directly restrict *alx1* and drive its expression. A positive regulatory input from *ets1* and, beyond 14hpf, by *tgif* has been proposed and supported by morpholino knockdown (MO) in *S.purpuratus* (Oliveri et al., 2008). While these relationships are likely to be significant, up till now there have been no studies showing direct linkages between these inputs and *alx1* expression. Furthermore, these inputs alone are insufficient to explain *alx1*'s distinctive mRNA kinetics. In Chapter 2 I will present evidence describing the causal regulatory inputs required for alx1 expression.

Sp-tbrain (*tbr*) is a member of family of transcription factors that contain a conserved DNA-binding domain called the T-box. T-box-encoding genes, of which brachyury is the founder member, have been shown to play important roles in both development of vertebrate and invertebrate endomesoderm, i.e. in gastrulation and heart development (Papaioannou and Silver, 1998). *Tbr* is maternally expressed but its zygotic transcription is restricted to the large micromeres at 7th cleavage and remains in their descendants throughout late larval stages. Its expression is driven by a ubiquitous maternal *ets1* input and constrained to the large micromeres by the double negative gate of *pmar1* and *hesc* (Wahl et al., 2009). Tbrain plays essential roles in both early specification and late differentiation. By activating the Ets factor *erg*, *tbr* switches on a positive feedback circuit involving *erg, hex, tgif* and *alx1* that locks down the skeletogenic regulatory state. *Tbrain* also collaborates with other transcription factors to turn on the differentiation battery required for early stages in spiculogenesis as well as postgastrular formation of the larval spicules (Fuchikami et al., 2002). The use of *tbrain* exclusively for sea urchin skeletogenesis is, from an evolutionary perspective, an interesting example of cooption of regulatory genes because its plesiomorphic expression pattern appears to be in the endomesoderm as evidenced in the sea cucumber and sea star embryos (Hinman and Davidson, 2007; Maruyama, 2000).

Among the four early regulators of skeletogenesis, the cis-regulatory architectures of the ets genes, *ets1* and *tel,* are the least understood. The ets family of transcription factors is widely used among metazoans. Ets factors play roles in cell proliferation, apoptosis, differentiation, migration and hematopoiesis (Sharrocks, 2001) and are characterized by the presence of a highly conserved DNA-binding domain known as the ETS domain, a tendency to interact with co-regulatory partner proteins and the ability to be a target of the MAP kinase signal cascade. The ETS domain of sea urchin *ets1/2* is most closely related to mammalian Ets1 and Ets2 and drosophila *pointed1* (PNT1), whereas *tel* is most similar to drosophila *yan* (Rizzo et al., 2006). In the sea urchin, *ets1/2* is required for specification of all mesodermal cell types. Like *tbrain*, *ets1/2* is both maternally and zygotically expressed. Zygotic expression begins downstream of the double negative gate, and remains exclusively in the SM lineage until late mesenchyme blastula stage. Expression then expands to the NSM lineage and remains in both lineages through late gastrulation. In the large micromeres, *ets1* is required to drive both *alx1* and *tbrain*. As such, it has a hand in all phases of skeletogenesis including the stabilization of the skeletogenic regulatory state, ingression and differentiation. Ets also plays *alx1* and *tbrain*-independent roles in the upregulation of the erg/hex/tgif, the expression of vegf-receptor, which is necessary for receiving signals that direct spatial organization of the skeleton, and expression of a subset of differentiation genes including Sm50. The observation that early specification regulators *ets1* and *alx1* participate in both the expression of differentiation drivers and the terminal gene battery itself is now a common one in network biology.

Subcircuitry for the stabilization of skeletogenic regulatory state

The genes *hex, erg* and *tgif,* as described earlier, form a subcircuit that acts as a dynamic stabilization device. Erg is a member of the ets family of transcription factors (Rizzo et al., 2006), and both hex and tgif are homeobox factors (Howard-Ashby et al., 2006b). Circuitry of this form is now understood to be a common feature of specification regulatory networks. They function to translate the transient transcriptional inputs that initialize a specification state into a persistent regulatory state (Davidson, 2010; Hinman et al., 2003). These circuits are able to stabilize through extensive positive cross-regulatory feedback as is seen between Hex and Tgif as well as from hex and erg and from Erg to hex and tgif. This feedback circuit is triggered by tbrain and ets as described above, and performs two important stabilizing functions: 1) tgif positively regulates alx1, driving its expression at early mesenchyme blastula stage and 2) Hex and Erg are positive drivers of the skeletogenic differentiation battery. Interestingly, the stabilization of skeletogenesis is disassociated from ingression, which is controlled primarily by alx1 and ets1 expression.

Signaling and regulators of morphogenesis

EMT transition of sea urchin mesenchymal is controlled by transcription factors that regulate the expression of mesenchyme-specific cadherins (N-cadherins) and other adhesion proteins as well as induce changes in cellular structure. In *Lythechinus*, the regulator *twist* has been implicated in directing this role. It activates *lv-snail* and permits the continued expression of *alx1* at mesenchyme blastula stage. Twist knockdowns don't appear to block ingression of PMCs, they just delay ingression and disrupt the pattern of migration. We can think of twist therefore as a global competence factor for mesoderm (Wu et al., 2008). *Lv-twist* is also required in the process of pigment cell specification as well as specification of muscle cells. In *S.purpuratus* the expression and function of *twist* has yet to be elucidated and *snail* expression was not detected in the large micromeres prior to ingression (S. Materna, unpublished).

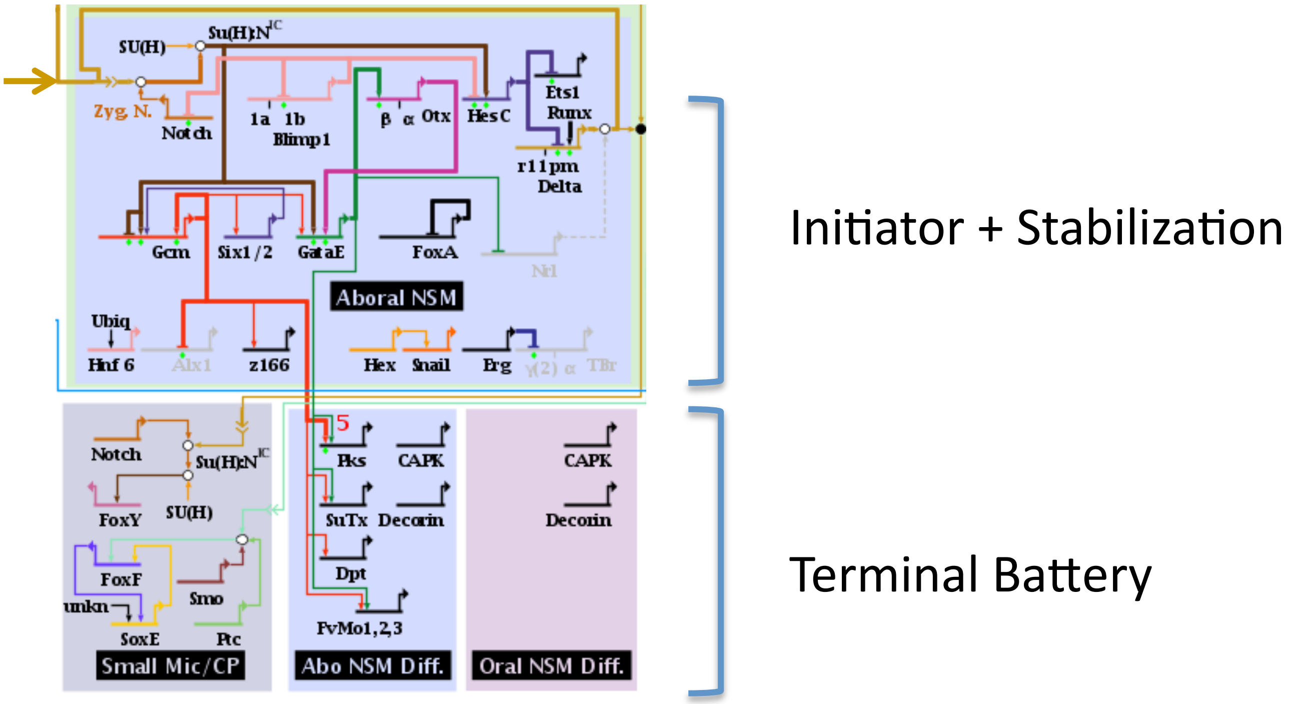
Expression of *vegf receptor* and fgf *receptor* permit the SM lineage to respond to VegF and FGF signals emitted by ventrolateral ectoderm at gastrula stages and beyond.

These signals cause a subpopulation of PMCs to aggregate in the adjoining blastocoel into two ventrolateral clusters that then nucleate the deposition of skeleton. The spatial position of these organizing signals is triangulated by repression from the Nodal signal of the oral ectoderm and positive inputs from adjacent Bmp2/4 signaling. Fgf is more important for later stages of skeletogenesis. FgfA signals emanate from ectoderm as a consequence of Nodal signaling and act by inducing sm30 and sm50 expression in the skeletogenic mesenchyme. These signals, once received, also guide migration of SM cells and control skeletal morphogenesis.

NONSKELETOGENIC MESENCHYME

The NSM lineage originates from veg2 blastomeres of the 7th cleavage embryo that are physically adjacent to the Delta-expressing large micromeres. While we now know the identity of many of the regulators expressed in the NSM lineage, their roles in the NSM GRN have yet to be elucidated. Also, later steps in NSM specification have been so far difficult to characterize due to a lack of sufficient early molecular markers for the various types of mesodermal tissue. The gene-regulatory network architecture of the NSM pathway, in regards to a subset of NSMs, the pigment cells, is fairly well-understood and is reflective of Type I embryogenesis (Davidson, 1991) (Figure 0.2a). Namely, it consists of a shallow network structure, containing a small number of transcriptional regulators that act both as specification factors that address a particular cell type as well as direct drivers of the differentiation battery required by that cell type.

This kind of shallow network hierarchy requires that these dual-purpose transcription factors contain cis-regulatory modules that interpret a wide variety of spatial and temporal queues, both repression and activation. An example of this is seen in the gcm cis-regulatory architecture as we will see below.



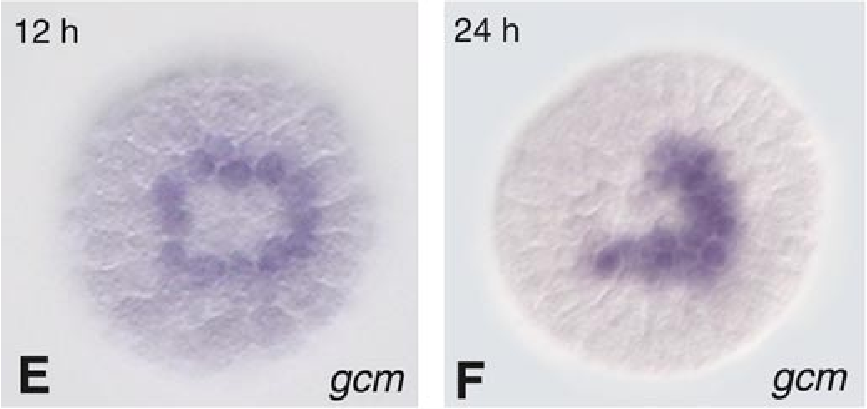


Figure 0.2. The pigment cell lineage. (A) Gene regulatory network for specifying pigment cell lineage. Delta signaling from neighboring large micromeres drives expression of N targets, including *gcm* and other transcription factors required for NSM specification. (B) Expression pattern for *gcm,* the terminal differentiation regulator of pigment differentiation. *Gcm* is initially expressed in all precursors of the NSM lineage and is subsequently restricted to the aboral mesoderm, which give rise to pigment cells (image from Ransick etal. 2002)

Regulators of NSM specification

Gcm is the earliest gene expressed downstream of delta-notch signaling in the veg2 lineage. Its functions are described here. Other genes expressed early in NSM specification are *gata-e*, *gata-c*, *ese, prox* and *scl* (S. Materna, A. Ransick, J. Rast and others, unpublished). *Spgcm* is the ortholog of the drosophila *glial cells missing* gene. Sea urchin genome sequencing has revealed the sea urchin has only 1 copy of *gcm* whereas Drosophila and human both have 2 copies. In Drosophila, *gcm* acts as a regulator of terminal differentiation in glial cells and in the specification and differentiation of macrophages. The roles for *gcm* in Drosophila hematocyte specification show interesting parallels to those seen in sea urchin NSM specification. In the fly embryo, blood cell precursors give rise to two cell types, precursors of plasmatocytes and crystal cells. *Gcm* expression is among the earliest regulatory switches for plasmatocyte specification, whereas the Runt-family transcription factor *Lozenge* (*Lz*) is required for crystal cell specification. Overexpression of *gcm* in crystal cell precursors is sufficient to convert these cells into plasmatocytes (Alfonso and Jones, 2002; Lebestky et al., 2000), revealing *gcm* plays roles in both promoting cell fate and also cross-repression of alternate similar fates. *Gcm* and its ortholog, *gcm2,* also behave downstream of plasmatocyte specification, as terminal regulators of macrophage differentiation, and knockout of both regulators causes improper migration, loss of expression of terminal differentiation genes such as Croquemort, and failure of plasmatocytes to convert to macrophages (Alfonso and Jones, 2002).

In the sea urchin, *gcm* is expressed directly downstream of the delta-notch signal, and functional Notch/SuH-interaction sites have been identified in its cis-regulatory genomic DNA (Ransick and Davidson, 2006). *Gcm* is expressed first in the 16 macromeres adjacent to the large micrometers at 7th cleavage (around 12-15 hpf) (Figure 0.2b) These cells constitute the Veg2 lineage and a subset of their descendants make up the entire nonskeletogenic mesoderm lineage. After initially expressing *gcm*, the veg2 tier cells undergo an additional division along the animal-vegetal axis, creating two distinct tiers of cells, namely those that continue to be adjacent to the large-micromeres (the veg2L, veg2 lower tier) and therefore continue to receive the delta/notch signal and express *gcm*, and those cells that are no longer in contact with Delta (the veg2U, or veg2 upper tier). The veg2L tier becomes the NSM lineage. Afterwards, *gcm* is downregulated in the oral quadrant of the NSM tier just prior to mesenchyme blastula stage (around 18-20 hpf), but transcripts remain in the aboral portion throughout gastrulation. This aboral subset of veg2L cells ingress into the blastocoel during gastrulation and migrate to the aboral ectoderm, where they differentiate into pigment cells. Careful cell counts and vital dye staining of NSM lineages showed that all of the gcm positive cells of the aboral mesoderm adopt a pigment cell fate and *gcm* transcripts can continue to be detected in the differentiated pigment cells of the aboral ectoderm. Pigment cells are so named for the presence of clusters of dark pigment granules that they contain. Their function in embryogenesis is not fully understood, although they have a cell morphology including irregular shape and multiple pseudopodia that is reminiscent of vertebrate macrophage cells. Also, experiments by J. Rast (unpublished) have shown that pigmented cells of the aboral ectoderm are capable of ingression into the blastocoel and phagocytosis of bacteria. These findings show that in the sea urchin embryo as in the fruit fly, *gcm* plays roles in the development of macrophage-like, innate embryonic immune cells.

The differences between its early and late spatial expression pattern suggest *SpGcm* may play two roles in mesoderm specification. Its early expression in all precursors of NSM show that the *gcm* cis-regulatory architecture is capable of interpreting the initial inputs and N signaling that specifies mesoderm. *Gcm* expression here creates a cell state that is competent to respond to later cues that direct development of the many mesodermal sub-lineages. Some evidence for this idea already exists. For example, while morpholino knockdown of *gcm* does not appreciably change the level of *gata-c* expression at midblastula stage, the spatial pattern of expression within the oral mesoderm becomes disorganized (A. Ransick, unpublished data). Also, morpholino knockdown of gcm expression leads to a loss of pigment cells, but also a yet-unexplained failure in the embryo’s ability to ingest food particles, a phenotype which may be attributed to improper development of the smooth muscle lining the sphincters of the tripartite gut. The later role of *SpGcm* in pigment cell specification is more straightfoward. Embryos injected with morpholino against *SpGcm* do not express the late pigment cell markers *SpPks*, *SpFMO* and *SpSULT* (A Ransick, unpublished data). A cis-regulatory analysis of the *pks* gene identified functional gcm-binding sites within 2kb upstream of the *pks* transcription start (Calestani and Rogers, 2010). Hence *gcm* acts here as a terminal differentiation regulator that drives several of the genes that build the unique morphology of the pigment cell.

When at 18-20hpf *gcm* expression retreats from the oral side it is exactly replaced by *SpGatac* expression. GataC is an ortholog of gata1/2/3 (Pancer et al., 1999) and is ultimately expressed in the blastocoelar subset of NSM cells of the embryo and in adult coelomocytes (Pancer et al., 1999). Injection of morpholino against *SpGatac* has no effect on *SpGcm* expression, however this perturbation induces expression of pigment-specific markers in the entire NSM lineage, suggesting GataC normally plays a role in blocking pigment cell fate in oral mesodermal cells (A. Ransick, unpublished data). The roles of the remaining known oral NSM-specific transcription factors *ese, prox* and *scl* are currently being elucidated.

Lockdown of Pigment Cell Fate

The regulatory network responsible for stabilizing the pigment cell fate is now being understood (A. Ransick and S. Materna) and appears to be controlled in a large part by the cis-regulatory architecture of *gcm* and by expression of a transcriptional activator, *six1/2. Gcm* expression is initiated by an early module that responds to Delta/Notch signaling originating from the large micromeres and acting through Supressor of Hairless (SuH) sites near the *gcm* locus. After SM cells ingress, however, NSM precursors are no longer in contact with Delta and therefore require an alternate mechanism for maintaining *gcm.* This role is performed by a second, late cis-regulatory module that contains binding sites for *gcm* itself, and for a co-activator, *six1/2,* whose expression is itself downstream of *gcm*. Hence, the maintenance subcircuit for pigment-cell specification involves a positive autoregulatory feedback from *gcm* and a second coherent positive input from *six1/2*.

Signaling to the NSM

Nodal and bmp2/4 signaling represent important inputs for patterning mesoderm. These signaling genes belong to the TGF beta superfamily of ligands. In TGF beta signaling, the binding of ligands to membrane-bound TGF-beta type II receptors induces phosphorylation of type I receptors which then lead to the phosphorylation of transcription factors called SMADs that translocate to the nucleus drive gene expression. The sea urchin *nodal* ortholog was first identified as an essential regulator in the patterning of ectoderm along the oral-aboral axis (Duboc et al., 2004), but additional studies by Duboc (Duboc et al., 2010) showed that nodal expression in the oral hemisphere polarized endoderm and mesoderm cell types along the oral-aboral axis as well. In the sea urchin, nodal signaling in the oral ectoderm induces bmp2/4 expression and signaling that extends further aborally. This splits the oral hemisphere into two regions, an oral pole that expresses both *nodal* and *bmp2/4* and an outer band that expresses only *bmp2/4*. The activities of *nodal* and *bmp2/4* are known to be antagonistic. Nodal promotes blastocoealar mesoderm specification through the alk4/5/7 receptor, which is expressed in these cells, and which is required here for blocking *gcm* expression in the oral-mesoderm (Duboc et al., 2010). The same receptor is required to turn on gata-c expression in those cells. BMP2/4 on the other hand, represses gata-c in a manner which appears opposite from its relationship to gata factors in vertebrates. Bmp2/4 signaling is required for driving gcm expression in the aboral ectoderm, and morpholino knockdowns of bmp2/4 cause an albino (pigmentless phenotype).

Hedgehog signaling also affects patterning and differentiation of NSM and SM descendants. At gastrulation, hedgehog is expressed in the endoderm and regulated by the Brachyury and Foxa transcription factors. This signal is received by the coreceptors Patched and smoothened (Ptc and Smo) which are expressed in neighboring mesenchyme. Perturbation of this signal doesn't effect specification of early mesoderm, but instead affects patterning of the differentiated mesenchymal cell types, and results in the alteration of numbers of pigment and blastocoelar cells, changes to left-right asymmetry in the coelomic pouches, and disorganization of esophageal muscle and derangement of the skeleton patterning (Walton et al., 2009).

CROSS-REPRESSION AND "FAIL-SAFE" WIRING IN MESODERM SPECIFICATION

Gene regulatory network analysis has revealed how the deployment of specification pathways is robust to small perturbations. It achieves this by a series of cross-repressive and redundant wiring systems whose job is to ensure that similar cell types do not accidentally adopt one another's specification states. We have already seen an example of this kind of wiring in the double-negative regulatory gate discussed earlier. Now I will briefly review additional examples that pertain to mesoderm specification.

The large micromeres and veg2L tier of cells share, in some ways, a similar regulatory state. The descendants of both lineages eventually express a common cohort of transcription factors including *ets1/2, erg, hex,* and both undergo similar EMT transitions and activate molecular machinery required for cell migration. These cells also have a similar developmental potential. NSMs have been shown to transfate SM at the expense of pigment and blastocoelar cell fates (Ettensohn and Ruffins, 1993) in micromereless embryos. As a consequence, the regulatory genome encodes multiple mechanisms for exclusion of alternative fates. In NSM specification, Delta/Notch signaling acts as a two-state switch. The target of canonical N signaling is the transcription factor suppressor of hairless Su(H)/CSL which, in the absence of signaling, recruits a co-repressor complex that shuts down transcription of a number of genes. The reception of a delta signal induces the cleavage of Notch protein and nuclearization of its intracellular domain which binds Su(H)/CSL and replaces the co-repressor with a co-activator complex that drives gene expression. In this manner, the Delta/Notch systems can lock out NSM specification in all cells but those that directly abut the large micromeres, namely the veg2L tier. Other direct forms of cell fate exclusion also exist. The expression of *alx1* in the large micromeres seems to prevent *gcm* expression in the SM lineage, although the mechanism for this is not clear (Oliveri et al., 2008). And as we will see in Chapters 1 and 2, *gcm* also has the capacity to shut down *alx1* in an *ets1*-dependent manner. Also, *tbrain* expression is blocked from being expressed in the NSM lineage by an erg-dependent cis-regulatory module that prevents ets1 from driving expression (Wahl et al., 2009). In a similar manner, late endoderm and NSM lineages must also share common regulatory inputs due to a requirement of *foxa* for blocking *gcm* in the developing archenteron (Oliveri et al., 2008).

Recently a type of redundant fail-safe wiring was observed for the micromere-specific repression of *hesc* (Smith and Davidson, 2009). It was noticed that *pmar1* knockdown did not prevent but only delayed specification of the large micromeres. This led to the discovery that *blimp1b*, which responds to the same inputs (*otx* and *tcf*) that drive *pmar1* acts as a delayed repressor of *hesc.* This regulatory wiring exists primarily as part of dynamic circuit driven by *wnt* and leading to a wave of beta-catenin nuclearization that starts at the vegetal pole and extends throughout the vegetal hemisphere in an expanding torus. The more ancient function of the *blimp1b* repression circuit may be to clear *hesc* so that its canonical repression target, *delta* can be expressed, first in the large micromeres and later in the presumptive NSM. However, because *hesc* repression is also connected to skeletogenic specification, the existing *blimp1b* wiring offers this second independent mechanism to *pmar1* expression for ensuring correct specification.

An additional layer of wiring discovered by Smith and Davidson (Smith and Davidson, 2009) showed *pmar1* and *hesc* operate in a reciprocal repressive embrace. This particular reciprocal repression does not function as a “bidirectional switch” as is common among these forms of wiring. Instead it works unidirectionally, where the initial expression of pmar1 in the large micromeres, induced by the positive, micromere-specific inputs of nuclear otx and nuclear beta-catenin, prevents transcription of hesC in these cells. At the same time in the rest of the embryo, the absence of pmar1 allows hesc message to accumulate. Later on in development, when nuclear otx and beta-catenin become abundant in other parts of the embryo, the presence of HesC protein prevents *pmar1* transcription, and ultimately blocks skeletogenesis. In this manner, the double negative gate works exclusively in the micromere lineage, as it should.

CONCLUSION

The study of the structure of developmental gene regulatory networks has revealed common themes about how specification occurs. The most basic specification network is composed of three layers. The first layer is an initiation subcircuit that integrates information from broadly expressed, and often partially overlapping spatial cues to direct regional expression of regulatory factors. Because initial cues are typically transiently expressed, a second "commitment" layer is necessary to lock down cell fate. This second layer is triggered by regulatory factors expressed in the initiation subcircuit, but positive cross-regulatory wiring of regulatory factors in this layer allow their expression to continue independently of the initiation layer circuitry. Often is the case, as seen in the SM lineage with the feedback of *tgif* onto *alx1,* that the maintenance layer stabilizes expression of initiation factors. In addition to lockdown of regulatory fate, both the initiation and maintenance circuitry often play a role in cross-repression of similar alternate cell fates. These and other redundant and failsafe design features ensure that developmental programs are executed consistently and explain how development proceeds unidirectionally. Finally, regulatory factors expressed in both layers turn on the gene batteries required for cell differentiation.

SUMMARY

In this Introduction I have reviewed the regulatory pathways involved in specifying embryonic mesoderm. In the following chapters I will attempt to fill in the regulatory logic governing a subsection of the mesoderm GRN and present new tools for authenticating gene regulatory networks and performing cis-regulatory analysis. In Chapter 1 I will discuss a new approach for testing developmental predictions made by the GRN that involves reengineering networks at the genomic level in vivo and studying developmental outcomes. In the process I will reveal how this approach can also be a useful tool for identifying hidden cross-regulatory wiring. In Chapter 2 I present a cis-regulatory analysis of the early co-regulator of skeletogenesis *alx1* that will reveal how genomic regulatory wiring precisely controls its temporal kinetics and spatial expression pattern. Chapter 3 will present a novel method for doing cis-regulatory analysis that employs quantitative imaging of reporter levels in live sea urchin embryos. Finally, in the appendix I describe a webtool written for the cis-regulatory biologist that is designed for fast annotation of short sequences (under 200kb) using binding site database information and sequence comparison algorithms.

REFERENCES

Alfonso, T.B., Jones, B.W., 2002. gcm2 promotes glial cell differentiation and is required with glial cells missing for macrophage development in Drosophila. Dev Biol 248, 369-383.

Calestani, C., Rogers, D.J., 2010. Cis-regulatory analysis of the sea urchin pigment cell gene polyketide synthase. Developmental Biology 340, 249-255.

Chuang, C.K., Wikramanayake, A.H., Mao, C.A., Li, X., 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, 231-237.

Davidson, E.H., 1991. Spatial mechanisms of gene regulation in metazoan embryos. Development 113, 1-26.

Davidson, E.H., 2010. Emerging properties of animal gene regulatory networks. Nature 468, 911-920.

Davidson, E.H., Cameron, R.A., Ransick, A., 1998. Specification of cell fate in the sea urchin embryo: summary and some proposed mechanisms. Development 125, 3269-3290.

Davidson, E.H., Rast, J.P., Oliveri, P., Ransick, A., Calestani, C., Yuh, C.-H., Minokawa, T., Amore, G., Hinman, V., Arenas-Mena, C., Otim, O., Brown, C.T., Livi, C.B., Lee, P.Y., Revilla, R., Schilstra, M.J., Clarke, P.J.C., Rust, A.G., Pan, Z., Arnone, M.I., Rowen, L., Cameron, R.A., McClay, D.R., Hood, L., Bolouri, H., 2002. A provisional regulatory gene network for specification of endomesoderm in the sea urchin embryo. Developmental Biology 246, 162-190.

Duboc, V., Lapraz, F., Saudemont, A., Bessodes, N., Mekpoh, F., Haillot, E., Quirin, M., Lepage, T., 2010. Nodal and BMP2/4 pattern the mesoderm and endoderm during development of the sea urchin embryo. Development 137, 223-235.

Duboc, V., Röttinger, E., Besnardeau, L., Lepage, T., 2004. Nodal and BMP2/4 signaling organizes the oral-aboral axis of the sea urchin embryo. Dev Cell 6, 397-410.

Ettensohn, C.A., Illies, M.R., Oliveri, P., 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., Ruffins, S.W., 1993. Mesodermal cell interactions in the sea urchin embryo: properties of skeletogenic secondary mesenchyme cells. Development 117, 1275-1285.

Fuchikami, T., Mitsunaga-Nakatsubo, K., Amemiya, S., Hosomi, T., Watanabe, T., Kurokawa, D., Kataoka, M., Harada, Y., Satoh, N., Kusunoki, S., Takata, K., Shimotori, T., Yamamoto, T., Sakamoto, N., Shimada, H., Akasaka, K., 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.

Hinman, V.F., Davidson, E.H., 2007. Evolutionary plasticity of developmental gene regulatory network architecture. Proc Natl Acad Sci USA 104, 19404-19409.

Hinman, V.F., Nguyen, A.T., Cameron, R.A., Davidson, E.H., 2003. Developmental gene regulatory network architecture across 500 million years of echinoderm evolution. Proc Natl Acad Sci USA 100, 13356-13361.

Howard-Ashby, M., Materna, S.C., Brown, C.T., Chen, L., Cameron, R.A., Davidson, E.H., 2006a. Gene families encoding transcription factors expressed in early development of Strongylocentrotus purpuratus. Dev Biol 300, 90-107.

Howard-Ashby, M., Materna, S.C., Brown, C.T., Chen, L., Cameron, R.A., Davidson, E.H., 2006b. Identification and characterization of homeobox transcription factor genes in Strongylocentrotus purpuratus, and their expression in embryonic development. Developmental Biology 300, 74-89.

Kenny, A.P., Kozlowski, D., Oleksyn, D.W., Angerer, L.M., Angerer, R.C., 1999. SpSoxB1, a maternally encoded transcription factor asymmetrically distributed among early sea urchin blastomeres. Development 126, 5473-5483.

Kumburegama, S., Wikramanayake, A.H., 2008. Wnt signaling in the early sea urchin embryo. Methods Mol Biol 469, 187-199.

Lebestky, T., Chang, T., Hartenstein, V., Banerjee, U., 2000. Specification of Drosophila hematopoietic lineage by conserved transcription factors. Science 288, 146-149.

Leonard, J.D., Ettensohn, C.A., 2007. Analysis of dishevelled localization and function in the early sea urchin embryo. Developmental Biology 306, 50-65.

Maruyama, Y.K., 2000. A Sea Cucumber Homolog of the Mouse T-Brain-1 is Expressed in the Invaginated Cells of the Early Gastrula in Holothuria leucospilota. Zool Sci 17, 383-387.

Materna, S.C., Howard-Ashby, M., Gray, R.F., Davidson, E.H., 2006. The C2H2 zinc finger genes of Strongylocentrotus purpuratus and their expression in embryonic development. Dev Biol 300, 108-120.

Okazaki, K., 1975. Spicule Formation by Isolated Micromeres of the Sea Urchin Embryo. American Zoology 15, 567-581.

Oliveri, P., Carrick, D.M., Davidson, E.H., 2002. A regulatory gene network that directs micromere specification in the sea urchin embryo. Developmental Biology 246, 209-228.

Oliveri, P., Tu, Q., Davidson, E.H., 2008. Global regulatory logic for specification of an embryonic cell lineage. Proc Natl Acad Sci USA 105, 5955-5962.

Pancer, Z., Rast, J.P., Davidson, E.H., 1999. Origins of immunity: transcription factors and homologues of effector genes of the vertebrate immune system expressed in sea urchin coelomocytes. Immunogenetics 49, 773-786.

Papaioannou, V.E., Silver, L.M., 1998. The T-box gene family. Bioessays 20, 9-19.

Peter, I.S., Davidson, E.H., 2009. Modularity and design principles in the sea urchin embryo gene regulatory network. FEBS Lett 583, 3948-3958.

Ransick, A., 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.

Revilla-i-Domingo, R., Oliveri, P., 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 USA 104, 12383-12388.

Rizzo, F., Fernandez-Serra, M., Squarzoni, P., Archimandritis, A., Arnone, M.I., 2006. Identification and developmental expression of the ets gene family in the sea urchin (Strongylocentrotus purpuratus). Developmental Biology 300, 35-48.

Sharrocks, A.D., 2001. The ETS-domain transcription factor family. Nat Rev Mol Cell Biol 2, 827-837.

Smith, J., Davidson, E.H., 2008. Gene regulatory network subcircuit controlling a dynamic spatial pattern of signaling in the sea urchin embryo. Proc Natl Acad Sci USA 105, 20089-20094.

Smith, J., Davidson, E.H., 2009. Regulative recovery in the sea urchin embryo and the stabilizing role of fail-safe gene network wiring. Proc Natl Acad Sci USA 106, 18291-18296.

Sodergren, E., Weinstock, G.M., Davidson, E.H., Cameron, R.A., Gibbs, R.A., Angerer, R.C., Angerer, L.M., Arnone, M.I., Burgess, D.R., Burke, R.D., Coffman, J.A., Dean, M., Elphick, M.R., Ettensohn, C.A., Foltz, K.R., Hamdoun, A., Hynes, R.O., Klein, W.H., Marzluff, W., McClay, D.R., Morris, R.L., Mushegian, A., Rast, J.P., Smith, L.C., Thorndyke, M.C., Vacquier, V.D., Wessel, G.M., Wray, G., Zhang, L., Elsik, C.G., Ermolaeva, O., Hlavina, W., Hofmann, G., Kitts, P., Landrum, M.J., Mackey, A.J., Maglott, D., Panopoulou, G., Poustka, A.J., Pruitt, K., Sapojnikov, V., Song, X., Souvorov, A., Solovyev, V., Wei, Z., Whittaker, C.A., Worley, K., Durbin, K.J., Shen, Y., Fedrigo, O., Garfield, D., Haygood, R., Primus, A., Satija, R., Severson, T., Gonzalez-Garay, M.L., Jackson, A.R., Milosavljevic, A., Tong, M., Killian, C.E., Livingston, B.T., Wilt, F.H., Adams, N., Belle, R., Carbonneau, S., Cheung, R., Cormier, P., Cosson, B., Croce, J., Fernandez-Guerra, A., Geneviere, A.M., Goel, M., Kelkar, H., Morales, J., Mulner-Lorillon, O., Robertson, A.J., Goldstone, J.V., Cole, B., Epel, D., Gold, B., Hahn, M.E., Howard-Ashby, M., Scally, M., Stegeman, J.J., Allgood, E.L., Cool, J., Judkins, K.M., McCafferty, S.S., Musante, A.M., Obar, R.A., Rawson, A.P., Rossetti, B.J., Gibbons, I.R., Hoffman, M.P., Leone, A., Istrail, S., Materna, S.C., Samanta, M.P., Stolc, V., Tongprasit, W., Tu, Q., Bergeron, K.F., Brandhorst, B.P., Whittle, J., Berney, K., Bottjer, D.J., Calestani, C., Peterson, K., Chow, E., Yuan, Q.A., Elhaik, E., Graur, D., Reese, J.T., Bosdet, I., Heesun, S., Marra, M.A., Schein, J., Anderson, M.K., Brockton, V., Buckley, K.M., Cohen, A.H., Fugmann, S.D., Hibino, T., Loza-Coll, M., Majeske, A.J., Messier, C., Nair, S.V., Pancer, Z., Terwilliger, D.P., Agca, C., Arboleda, E., Chen, N., Churcher, A.M., Hallbook, F., Humphrey, G.W., Idris, M.M., Kiyama, T., Liang, S., Mellott, D., Mu, X., Murray, G., Olinski, R.P., Raible, F., Rowe, M., Taylor, J.S., Tessmar-Raible, K., Wang, D., Wilson, K.H., Yaguchi, S., Gaasterland, T., Galindo, B.E., Gunaratne, H.J., Juliano, C., Kinukawa, M., Moy, G.W., Neill, A.T., Nomura, M., Raisch, M., Reade, A., Roux, M.M., Song, J.L., Su, Y.H., Townley, I.K., Voronina, E., Wong, J.L., Amore, G., Branno, M., Brown, E.R., Cavalieri, V., Duboc, V., Duloquin, L., Flytzanis, C., Gache, C., Lapraz, F., Lepage, T., Locascio, A., Martinez, P., Matassi, G., Matranga, V., Range, R., Rizzo, F., Rottinger, E., Beane, W., Bradham, C., Byrum, C., Glenn, T., Hussain, S., Manning, G., Miranda, E., Thomason, R., Walton, K., Wikramanayke, A., Wu, S.Y., Xu, R., Brown, C.T., Chen, L., Gray, R.F., Lee, P.Y., Nam, J., Oliveri, P., Smith, J., Muzny, D., Bell, S., Chacko, J., Cree, A., Curry, S., Davis, C., Dinh, H., Dugan-Rocha, S., Fowler, J., Gill, R., Hamilton, C., Hernandez, J., Hines, S., Hume, J., Jackson, L., Jolivet, A., Kovar, C., Lee, S., Lewis, L., Miner, G., Morgan, M., Nazareth, L.V., Okwuonu, G., Parker, D., Pu, L.L., Thorn, R., Wright, R., 2006. The genome of the sea urchin Strongylocentrotus purpuratus. Science 314, 941-952.

Su, Y.H., Li, E., Geiss, G.K., Longabaugh, W.J., Kramer, A., Davidson, E.H., 2009. A perturbation model of the gene regulatory network for oral and aboral ectoderm specification in the sea urchin embryo. Dev Biol 329, 410-421.

Wahl, M., Hahn, J., Gora, K., Davidson, E., 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. Developmental Biology.

Walton, K.D., Warner, J., Hertzler, P.H., McClay, D.R., 2009. Hedgehog signaling patterns mesoderm in the sea urchin. Developmental Biology 331, 26-37.

Wu, S.-Y., Yang, Y.-P., McClay, D.R., 2008. Twist is an essential regulator of the skeletogenic gene regulatory network in the sea urchin embryo. Developmental Biology 319, 406-415.

CHAPTER 1

An In Vivo Synthetic Approach to Network Validation and Discovery

Sagar Damle, Eric Davidson

(In preparation, *PNAS)*

Abstract

The gene regulatory networks controlling specification and development of skeletal mesenchyme (SM) in the purple sea urchin *S. purpuratus* embryo are now well understood. However an ultimate demonstration of a network's ability to explain the causal mechanisms that drive development will be to rewire its regulatory linkages in novel ways to produce predictable developmental outcomes. Here we use BAC recombinant constructs to introduce new regulatory wiring to the SM specification pathway. We bring *gcm,* a transcription factor that drives the differentiation battery for the pigment cell type under control of the double-negative regulatory gate responsible for specifying the skeletogenic lineage. We find that in the rewired regulatory state, *gcm* expression overrides the skeletogenic program in a way that validates some aspects of our network model and also reveals new cryptic repressive repression functions.

Keywords: sea urchin, mesoderm, gene regulatory network, re-engineering

INTRODUCTION

In gene regulatory networks (GRNs) controlling embryogenesis, transcription factors, and signaling molecules represent nodes whose interactions are integrated at the level of cis-regulatory DNA to drive development. The regulatory relationships in the sea urchin endomesoderm GRN in particular have been studied extensively for over a decade (Davidson et al., 2002; Peter and Davidson, 2009). Additionally, the recent completion of a draft genome sequence has made it possible to identify nearly all transcription factors and C2H2 zinc fingers in the sea urchin as well as characterize the spatial and temporal expression of those factors expressed during early embryogenesis from fertilization, through to late blastula and early gastrula (Howard-Ashby et al., 2006; Materna et al., 2006). However, an ultimate demonstration of our understanding of gene networks is to succeed in predicting the developmental consequences of re-organizing the network at the level of genomic DNA.

The sea urchin model system offers a unique combination of advantages for directed network testing. These include the following: i) embryos are easy to inject with DNA constructs, ii) embryos are transparent accessible to the imaging of multiple fluorescent reporters, iii) development, from cleavage to gastrulation, occurs within a span of 40 hours, iv) the regulatory networks responsible for endomesoderm specification are well understood and the ectodermal regulatory network is being actively studied and v) the sea urchin genome has been sequenced and the temporal and spatial expression pattern of a majority of transcription factors and signaling proteins has been catalogued. The idea of using gene transfer to alter morphogenetic or phenotypic outcomes is not new.

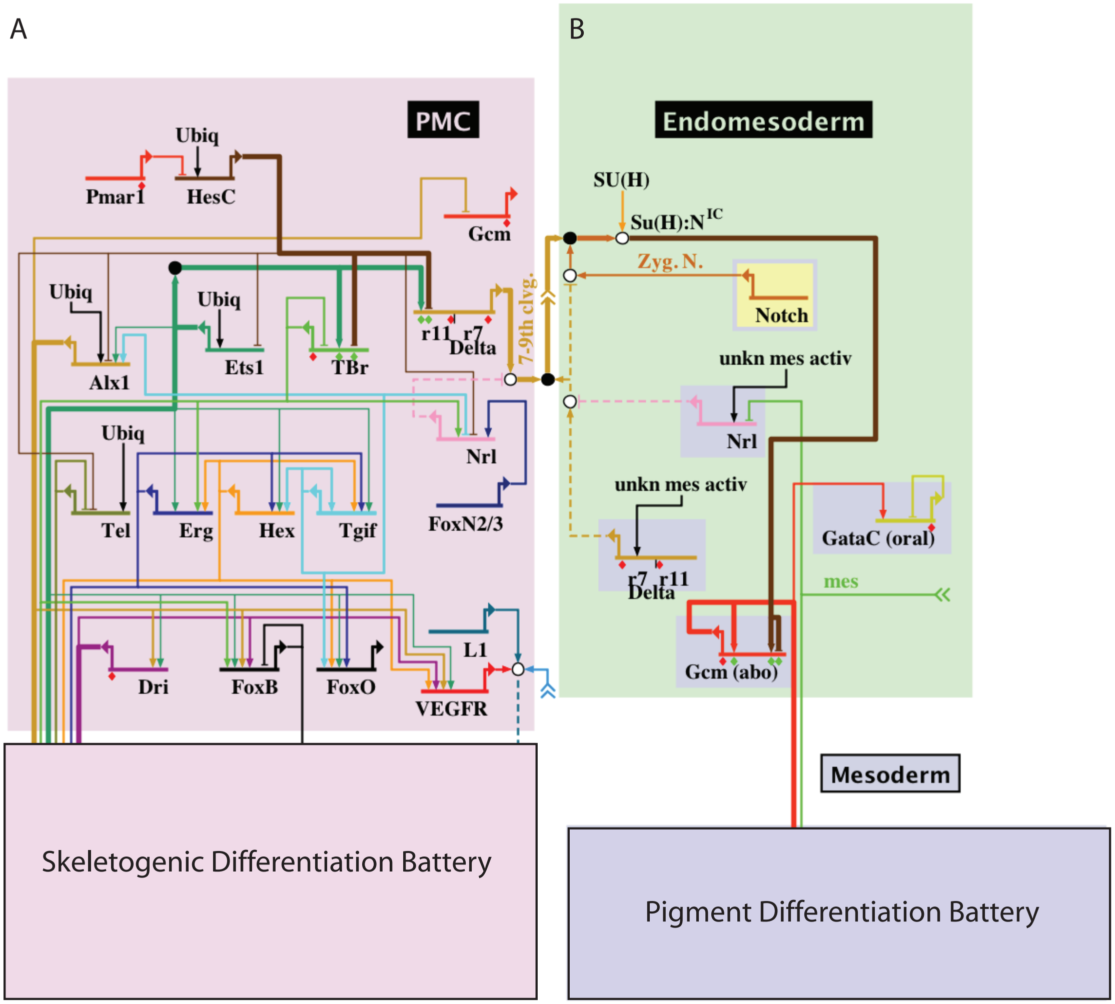
Misexpression of pax6 in both drosophila and vertebrate embryos has been used to demonstrate its sufficiency for activating the regulatory network responsible for lens and eye development (Altmann et al., 1997; Halder et al., 1995). More precise re-engineering of synthetic networks has even been demonstrated in bacterial systems (Elowitz and Leibler, 2000). This work is unique in that it combines spatial and kinetic precision to a developmental context to alter the specification network of a complex multicellular organism.

To that end, we have used BAC recombination (Lee et al., 2001) to generate a construct that expresses *gcm*, a critical regulatory factor for non-skeletogenic mesoderm specification and pigment cell differentiation under the control of the *tbrain* cis-regulatory architecture. Zygotic *tbrain* (*tbr*) expression begins in the large micromeres after 6th cleavage (roughly 8 hours post fertilization) and remains in the skeletogenic mesenchyme (SM) through late gastrula stage. When injected into developing embryos, this BAC construct drives *gcm* in SM precursors at 8-9th cleavage (early blastula stage) in a manner that resembles endogenous *tbrain* expression. The consequences of this rewiring are explored below.

RESULTS

The Mesoderm Specification GRN

The developing embryo initially gives rise to two types of mesodermal cells. At blastula stage, both types of cells ingress from the vegetal plate. The first cells to ingress, the so-called primary mesenchyme cells, develop into the embryonic skeleton, whereas secondary mesenchyme cells (or non skeletogenic mesenchyme, NSM) involute at the leading edge of the archenteron after primary ingression. Both cell types appear to express a similar cohort of transcription factors (ets1, delta, Hex) that precede or coincide with morphological changes associated with ingression and migration, and each type also expresses factors unique to the functions of their differing developmental fates. The SM have been hypothesized to be a specialized mesodermal cell type with a recent evolutionary history (Davidson and Erwin, 2006). And while not all echinoids have an embryonic skeleton, in *S. purpuratus* the regulatory network controlling adult skeletogenesis seems to have been coopted for embryogenesis, at a point just downstream of the double-negative gate, which is composed of the repressors *pmar1* and *hesc* (Gao and Davidson, 2008). A process diagram for specification of the SM lineage can be seen in Figure 1.1a and is summarized below. Maternal anisotropies at the 32 cell stage, including a bias towards Beta-catenin and Otx nuclearization at the vegetal pole, cause the expression of *pmar1* in the large micromeres. *Pmar1* acts as a repressor of the ubiquitously expressed *hesc* gene, which in turn represses a number of transcription factors necessary for initiating the skeletogenic program. *Hesc* is a direct target of tbrain (Wahl et al., 2009) and *alx1* (Damle, in press) and also regulates additional factors, *ets1/2* and *tel*, all of which are required to drive the skeletogenic differentiation battery. Also, *hesc* controls the expression of delta, which triggers NSM specification in neighboring macromeres, as we will describe next.



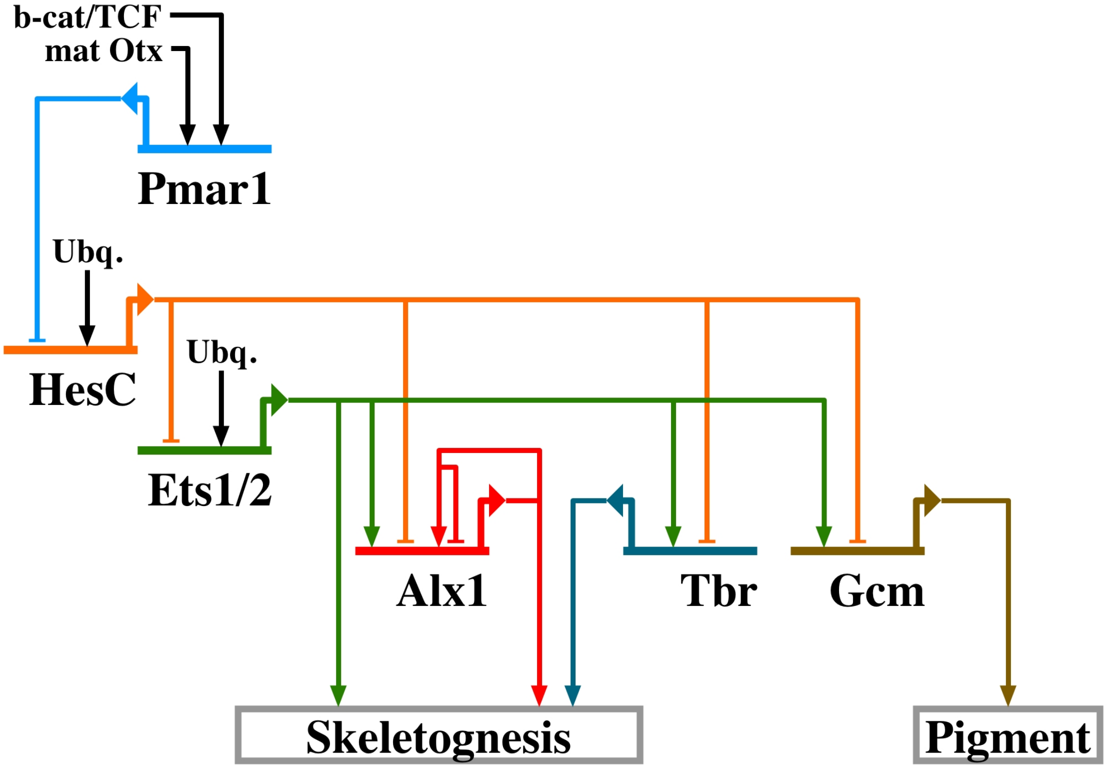


Figure 1.1. Endomesodermal gene-regulatory network showing PMC and mesoderm specification and differentiation circuits. (A) PMC specification begins with the expression of the repressor Pmar1. A double-repressive regulatory circuit leads to expression of the first tier of PMC regulatory genes: *alx1*, *ets1*, *tbrain*. These factors control timing of ingression into the blastocoel, and the morphological and transcriptional changes controlling skeletogenesis (B) The signaling factor Delta, expressed in large micromeres, signals through notch receptor to neighboring endomesoderm and induces Gcm expression in the progenitors of pigment cells. Gcm is necessary for expression of several genes involved in pigment biosynthesis and ingression. (C) Diagram of rewired skeletogenic specification network. Gcm expression is brought under control of the double-negative regulatory gate and is expressed in precursors to the skeletogenic mesenchyme. Given its capacity to autoactivate, synthetic gcm may also be capable of turning on the endogenous gcm regulatory wiring. Also, because gcm is a terminal regulator of the differentiation battery, it should activate the SMC/pigment cell differentiation network. If there are no cross-repressive functions of gcm on skeletogenic specification (shown by green repression bars on Alx1 and Ets1) then the skeletogenic regulatory network will proceed in parallel and unhindered by pigment differentiation.

Figure 1.1b describes the regulatory network architecture for NSM specification. Initiation is marked by the expression of *gcm* at 7th cleavage in endomesodermal precursors abutting the large micromeres, the veg2 tier of cells (Ransick et al., 2002). *Gcm* expression is induced by delta-notch signaling from adjacent large micromeres (Sweet et al., 2002; Sweet et al., 1999). The descendants of these *gcm*-positive cells give rise to all NSMs. After their initial expression, *gcm* transcripts become excluded from the oral quadrant of the veg2 tier. While both cell types ingress into the blastocoel, the *gcm*-negative oral quadrant develops into a wide variety of mesodermal cell types – including blastocoelar, muscle, and coelomic pouch cells – while the aboral *gcm*-positive mesenchyme differentiates into pigment cells, and migrates to and embeds in the aboral ectoderm. Embryos injected with morpholino antisense oligos against *gcm* fail to develop pigment cells, supporting its necessity for that developmental pathway. *Gcm* is a direct target of the pigment cell differentiation gene, polyketide synthase (pks) (Calestani and Rogers, 2010) and is capable of strongly upregulating the expression of a number of pigment-specific differentiation genes. These results identify *gcm* as a terminal differentiation transcription factor of the pigment cell type and predict that its expression should be sufficient to drive pigment cell differentiation in the NSM.

SM and NSM cell types share a similar developmental potential. NSM cells have been shown to be capable of rescuing skeletogenesis in embryos in which the ingressed SM cells have been removed. Similarly, SM lineage may also retain a potential for developing into NSM. Knockdown of *alx1*, which is required for skeletogenesis, causes *gcm* and *pks* expression in SM cells (Oliveri et al., 2008). In addition, all mesenchyme lineages in the sea urchin express a similar set of regulatory factors that are likely necessary for the common functions of all mesodermal cell types.

Given the similarities in these cell types, we decided to add regulatory wiring that would bring *gcm* expression under the control of the double-negative gate and cause it to be expressed in the large micromeres at a very early stage in SM specification (Figure 1.1c). We accomplished this by driving *gcm* expression using the *tbrain* cis-regulatory architecture. Croce et al. (2001) described the embryonic spatiotemporal expression pattern of the Tbrain orthologue, *ske-T*, in *Hemicentrotus pulcherrimus*. Like *tbrain, ske-T* is expressed maternally. Whole-mount in-situ hybridization (WMISH) experiments show zygotic *ske-T* expression occurs in primary mesenchyme cells at early blastula stage (prior to hatching) and persists through late gastrula in their descendants, the skeletogenic mesenchyme. More recently identical results have been obtained for *tbrain* in *S. purpuratus* (Wahl et al., 2009). These show an initiation of Tbrain expression at 10-14 hours, leading to peak expression at 21 hr (just prior to PMC ingression into the blastocoel). A second broad peak of expression has been observed at mid-gastrula (33-40 hpf).

This rewiring design removes the requirement for delta/notch signaling and therefore allows *gcm* to be expressed outside of cellular environments that contain repressive SuH/CSL complexes that directly inhibit endogenous *gcm.* This also disassociates *gcm* from any *alx1*-mediated repression pathways as *alx1* has no effect on the expression of *tbrain* in the SM lineage. Given our understanding of mesoderm specification there were two possible outcomes of this rewiring design: a) a complete coexpression of both SM and NSM differentiation pathways, and b) a dominance of pigment cell fate and simultaneous repression of skeletogenesis. While the current network model predicts that outcome (a) will occur, the observation of outcome (b) would imply the existence of additional exclusion functions for *gcm*.

Construction of recombinant BACs for synthetic rewiring

We added regulatory wiring to *gcm* expression through the use of BAC homologous recombination by inserting *gcm* coding sequence into the first exon of *tbrain*-containing BAC (Figure 1.2a). *Tbrain* is expressed as a consequence of the Pmar1-mediated double-negative gate (Wahl et al., 2009). This wiring permits *gcm* expression in a manner that accurately recapitulates the spatial and temporal expression pattern of *tbrain.* Embryos injected with the *tbr::gcm* BAC are capable of expressing in skeletogenic mesenchyme in large micromere clones as early as endogenous *tbr.*

In order to measure in vivo the regulatory state changes occurring in these *gcm*-expressing SM clones we generated a series of cell-state detector BACs and constructs (Figure 1.2b). We generated reporters for *alx1, ets1/2,* and *tbrain* expression by inserting GFP coding sequence into the first exon of BACs that contain each gene. We also measured the degree of pigment cell differentiation by generating an RFP construct that was driven by a cis-regulatory module that controls expression of the *pks* gene (Calestani and Rogers, 2010).

When injected into fertilized eggs, each BAC reporter and small construct was capable of recapitulating the expression pattern of its corresponding gene. Using qpcr we obtained precise measurements for the tbrain:GFP BAC expression at 15, 20, 30, and 43 hours post fertilization and compared them again endogenous *tbrain.* The expression data was normalized for number of copies of construct incorporated on average per embryo as described (Revilla-i-Domingo et al., 2004) and shows that the timing of initiation of GFP transcription matches that of the endogenous Tbrain (Supplemental Figure 1.1)

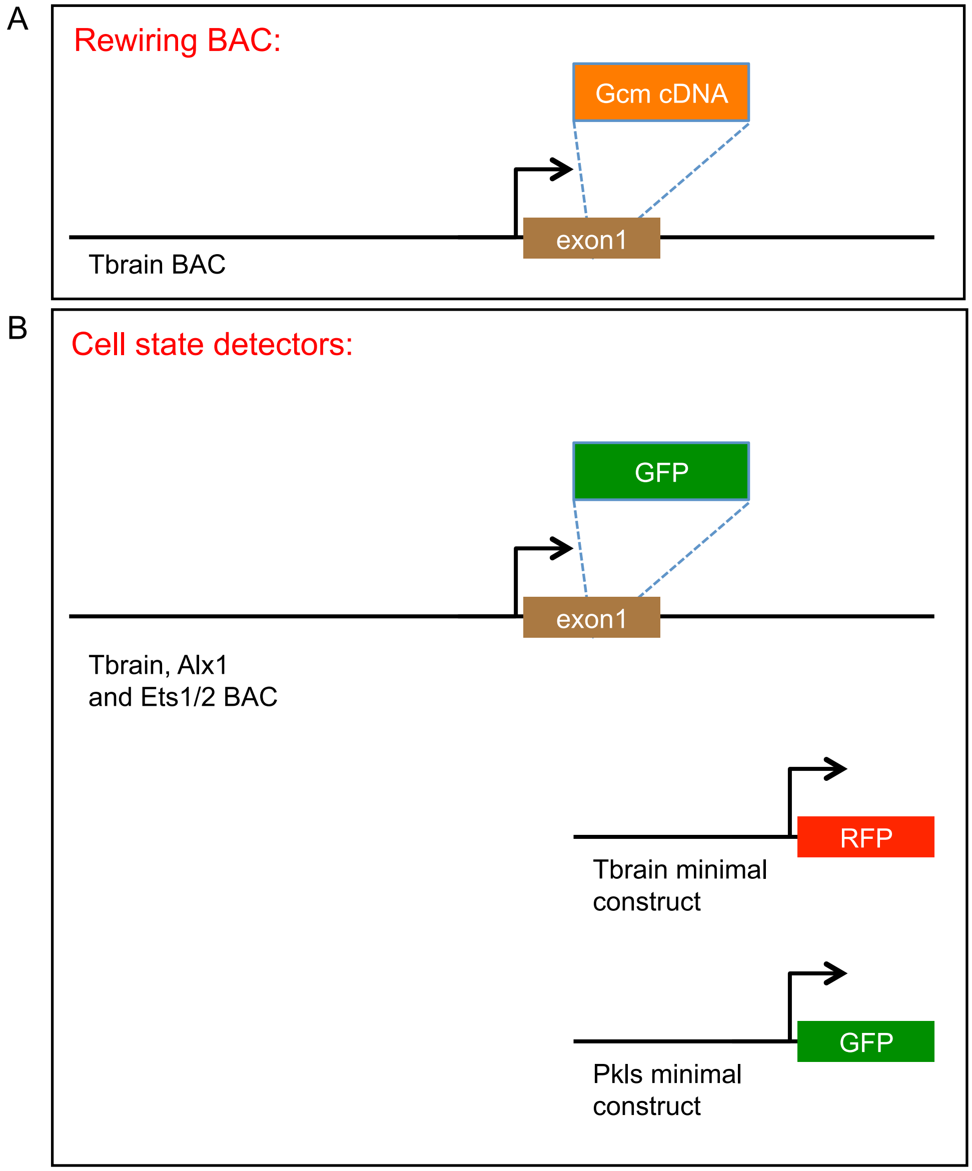


Figure 1.2. Diagram of BAC constructs used in rewiring experiment. A) *gcm* coding sequence was inserted using homologous recombination into the first exon of the *tbrain* gene in a 140 kb BAC that contains the entire tbrain regulatory architecture. B) A similar

knock-in strategy was used to generate BAC-GFP reporters that were used as detectors for measuring cell state. BAC-GFP constructs were made for the *tbrain, alx1,* and *ets1/2* gene. Short construct GFP reporters were made for detecting *tbrain* and *pks* expression. These constructs faithfully recapitulate the spatial expression patterns of their corresponding endogenous genes.

Next we observed GFP protein fluorescence in live animals over the course of early embryogenesis. We compared Tbrain-GFP BAC reporter expression to endogenous Tbrain expression during similar stages of development. Characteristic of many BAC reporters, the *Tbrain* GFP BAC shows precise temporal and spatial recapitulation of endogenous gene expression. Due to mosaic nuclear integration in developing embryos the reporter is initially expressed in only a fraction of SM cells. At late mesenchyme blastula stage however, SM fuse to form a syncytium – a necessary step in their construction of the embryonic skeleton. During this process, GFP protein distributes evenly such that every skeletogenic cell is fluorescent, as seen in Supplemental Figure 1.1, B4, although in situs performed on mRNA remain localized to the cells in which they are expressed (Damle, unpublished). Scoring of injected embryos shows the BAC is capable of matching the spatial Tbrain expression pattern, with very little ectopic expression (Supplemental Figure 1.1). Of expressing embryos, 80-95% show expression in the large micromeres and skeletogenic mesenchyme. An additional 5-10% show ectopic expression in blastocoelar cells. If followed late into development, it can be seen that no GFP+ cells express pigment. The results, taken together, confirm the BAC contains all the regulatory information necessary for recapitulating the spatiotemporal expression pattern of endogenous *tbrain*, and that there is no ectopic expression in pigment cells.

Fate transformation effects of synthetic gcm expression in skeletogenic cells

Two-color WMISH was performed on embryos injected with Tbrain:GCM BAC to look for pigment differentiation battery genes upregulated in the SM lineage (Figure 1.3a). These experiments showed that *pks* and *fmo* are upregulated in synthetic-*gcm*-expressing skeletal mesenchyme. In order to assay in real time the fate transformation in *gcm*-expressing SM cells, Tbrain-RFP and Pks-GFP small-construct reporters were generated such that their reporter expression matched that of endogenous *Tbrain* and *Pks* (Calestani and Rogers, 2010). *Pks* is a differentiation gene that is involved in pigment synthesis and expressed in the subset of SMC that eventually migrates to the aboral ectoderm. Hence these reporter constructs act as markers for either PMC or pigment cell types. These two were coinjected either in the presence or absence of a Tbrain-GCM BAC and observed during development. Figure 1.3b and Supplemental Figure 1.2 show that when TbrainRFP and PksGFP are coinjected, their expression patterns do not overlap in the developmental stages in which they are expressed. When coinjected in the absense of Tbrain-GCM, 6% of expressing embryos have cells that are both GFP and RFP positive. When coinjected in the presence of Tbrain-GCM, however, 30% of embryos contain PMCs that coexpress GFP and RFP (i.e., Pks and Tbrain positive) (Table 1.2, Figure 1.3b/c), indicating these cells have adopted a regulatory state that contains features unique to both SM and NSM lineages. It should be noted that a small fraction (9/161 or 5.5%) of RFP-positive, GFP-positive cells that otherwise behaved as SM cells remained in the syncytial skeletal rings.

Embryos injected with Tbrain::GCM BAC were cultured beyond last gastrula stage to assay effect of gcm activity on skeletogenesis. Table 1.1 shows that, of expressing embryos, 40-50% continue to show GFP expression in skeletogenic precursors, a 50% decrease in comparison to embryos injected with Tbrain:GFP alone. The remaining 40% of embryos show GFP positive cells in the blastocoel (roughly 30%) and in cells expressing pigment and residing in aboral ectoderm (15-20%). Figure 1.3D shows an uninjected embryo at 48hpf viewed form the vegetal pole. The PMCs can be seen arranged in a ring around the gut, outlining the location of the future growing arms of the skeleton. A representative embryo injected with Tbrain-BAC-GFP only at the same stage of development, showing a similar arrangement of syncytial PMCs and intact skeletal arm (Figure 1.3D). In embryos coinjected with Tbrain-GCM-BAC, GFP-positive SMs have failed to form a syncytium with their sister SMs, and also failed to arrange themselves in basal and lateral rings characteristic of normal SM. Instead, they exist in the blastocoel (arrow in Fig 1.3D) and in some cases migrate to the aboral ectoderm and even begin to express pigment. These results suggests respecification of SM occurs prior to syncytium formation as we will describe below.

Figure 3

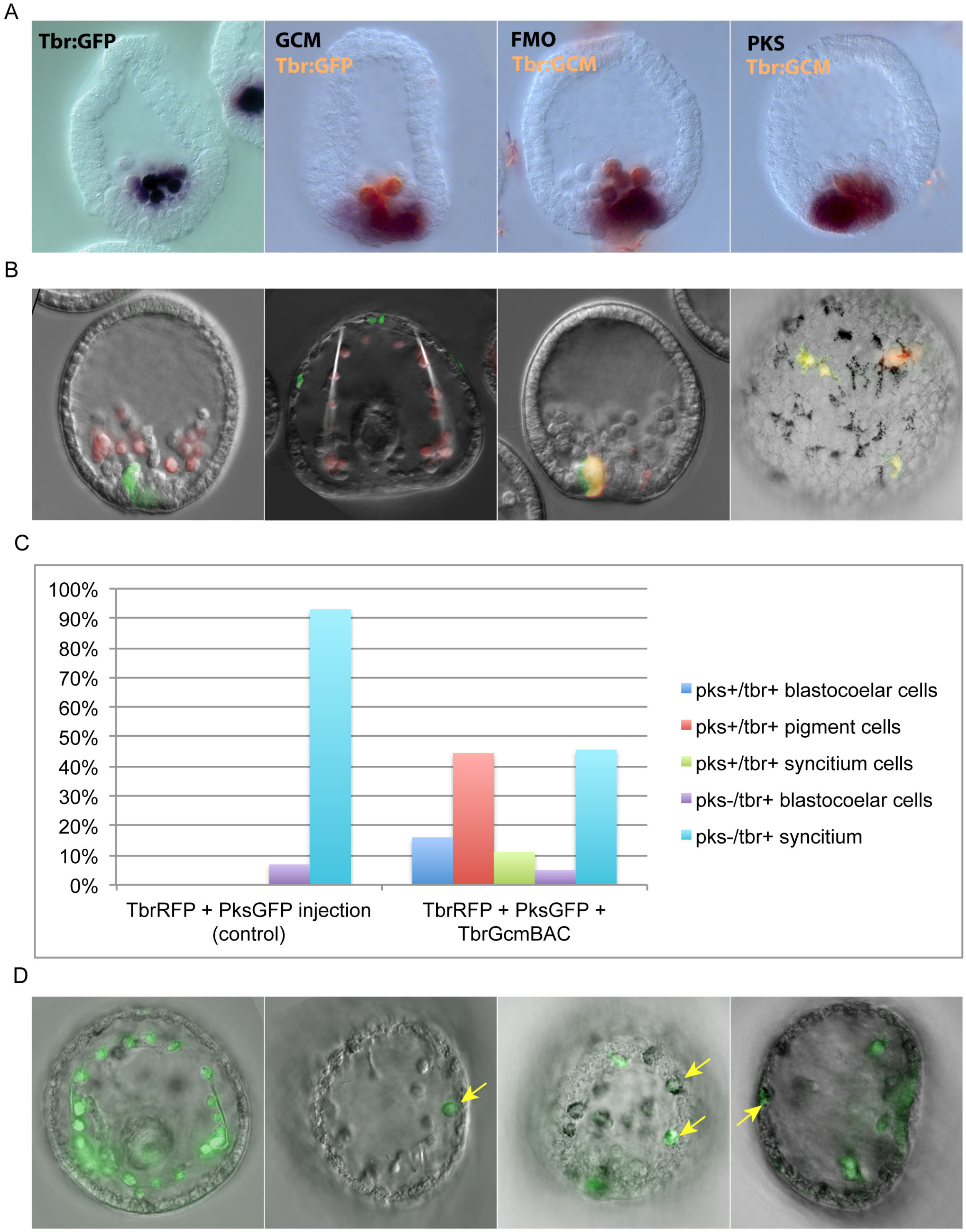

A
Tbr:GFP
GCM
Tbr:GFP
FMO
Tbr:GCM
PKS
Tbr:GCM
B
C
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
TbrRFP + PksGFP injection (control)
TbrRFP + PksGFP + TbrGcmBAC
pks+/tbr+ blastocoelar cells
pks+/tbr+ pigment cells
pks+/tbr+ syncitium cells
pks-/tbr+ blastocoelar cells
pks-/tbr+ syncitium
D

Figure 1.3. Fate transformation in *gcm*-expressing SM cells. A) 2-color WMISH of embryos at mesenchyme blastula stage. Probes used for detection are indicated at the top left of each image. In the case of Tbrain:GFP and Tbrain:GCM construct detection, the

probe used detected the 3'UTR SV40 poly-adenylation sequence. Both GFP and exogenous GCM contain identical SV40 3'UTRs. B) Tbrain:RFP and Pks:GFP detector coinjection experiments with Tbrain:GCM BAC. The images from left to right are as follows: detector coinjection without Tbrain:GCM BAC at i) mesenchyme blastula stage and at ii) late gastrula stage; detector coinjection *with* Tbrain:GCM BAC at iii) mesenchyme blastula stage and iv) late gastrula stage. D) Late differentiation of gcm-expressing SM cells. Left-most image is a control injection of Tbrain:GFP BAC only at late gastrula stage. Remaining 3 images show Tbrain:GFP and Tbrain:GCM BAC coinjections at late gastrula. Yellow arrows mark pigment granules in GFP-positive cells.

Synthetic expression reveals cryptic exclusion functions

Because the effect of *gcm* expression on SM specification appears to show incomplete dominance—there is a subpopulation of SM clones that express *gcm*, but nevertheless fuse to form a syncytium and contribute to the production of skeleton—it is difficult to use cell-sorting techniques to assay the effects on SM-specific gene markers in GFP-positive cell populations (data unpublished). To distinguish whether all or only a fraction of *gcm*-positive SMs show any repression of SM specific markers, 2-color whole-mount in-situ hybridization on injected embryos was performed. In these experiments, embryos coinjected with Tbrain::GCM were fixed at mid-blastula stage (22-24 hours post fertilization). Expression of exogenous *gcm* was observed by staining with a probe that matched its SV40 3'UTR sequence. Staining of similar-stage embryos injected with TbrainGFP construct showed nearly all embryos (80% of expressing samples) with strong expression of GFP in ingressing SM cells. Embryos injected with Tbrain::GCM, however, showed expression in cells at the base and periphery of the vegetal plate (65%), in addition to expression within ingressed SM cells (30%).

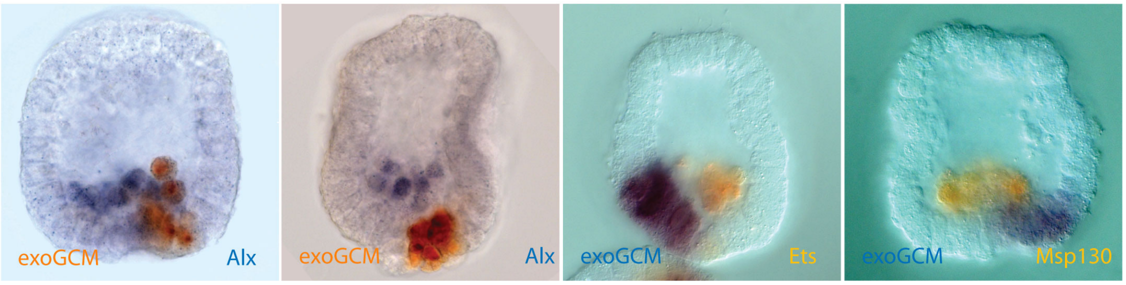


Figure 1.4. 2-color WMISH of gcm-expressing SM cells at mesenchyme blastula stage. Probes used are indicated at the bottom of each image.

Injected embryos were co-stained for the presence of a number of markers expressed in SM (in yellow). Figure 1.4 displays expression of various SM-specific markers in these embryos. In these animals, it is clear that the subset of GCM-positive SM cells that remain in the vegetal plate also show lack of expression of SM genes (ets1, msp130, jun, foxb) (See Figure 1.4 and data not shown) . On the other hand, as seen in Figure 1.4 (Alx expression in purple, and GCM expression in orange) the subset of *gcm*-positive SMs that have already ingressed into the blastocoel show continue to express *alx1*, whereas those SMs remaining in the vegetal plate show decreased or no Alx1. Thus, *gcm* expression in SM has been shown to be capable of repressing SM specification. We scanned for additional putative targets of *gcm* repression by injecting full-length *gcm* mRNA into fertilized eggs and performing qPCR at mesenchyme blastula stage of SM-specific genes (Supplemental Figure 1.5). This showed a sharp downregulation of *vegfrII*, *msp130L,* and *foxO.*

To confirm that a regulatory link between *gcm* and *alx1* exists, an Alx1::GFP reporter was generated and its activity was measured in the presence of *gcm* mRNA overexpression (*gcm* MOE). The minimal Alx1 reporter construct contains 400bp of genomic DNA located upstream of the start site of transcription and is capable of expressing GFP in SM precursors. *Gcm* MOE led to a 2-fold downregulation of GFP expression of the construct, however a *gcm*-responsive element was not found (data not shown). These truncations suggest the effect of *gcm* on *alx1* expression may be indirect and due to effects on upstream inputs to *alx1*.

An important regulatory input for *alx1* is the *ets1* gene. *Ets1* is maternally expressed (Rizzo et al., 2006) and remains ubiquitous through late cleavage stages. By early blastula stage *Ets1* becomes zygotically expressed in the skeletal precursors and precedes cell ingression. *Ets1* is similarly expressed in NSM cells at late mesenchyme blastula stage, also preceding their ingression. *Ets1* MASO injected embryos show downregulation of Alx1 at mesenchyme blastula stage (Davidson et al., 2002; Ettensohn et al., 2003) as well as downregulation of a number of SM-specific specification genes (*foxb, tgif, hex, erg, dri*) and components of the skeleton (*ficolin, sm50, cyclophilin*). Synthetic *gcm* expression downregulates *ets1* in the SM precursors that fail to ingress at mesenchyme blastula (24 hpf). While *Ets1* is expressed in both SM and NSM precursors, *alx1* is expressed uniquely in SM cells and is necessary for both ingression and skeletogenic differentiation. So it seems likely that *alx1* downregulation by *gcm* is mediated through *ets1* and leads to a subsequent failure in skeletogenesis.

DISCUSSION

Aspects of network wiring related to the structure of differentiation gene battery activating networks

The goal of this work were to test predictions made by the network about the sufficiency of the ability of certain critical nodes to drive specification programs in other contexts, and to perform these tests by rewiring in vivo the GRN. We have achieved this by bringing together the cis-regulatory DNA controlling expression of a SM-specific factor, *tbrain*, to the coding sequence for a gene responsible for pigment cell specification, *gcm* to produce a new mesodermal gene regulatory network in network in-vivo (figure 1.1C).

The GRN responsible for making pigment cells is relatively shallow. In the context of the undifferentiated mesodermal specification state, a single gene, *gcm* seems sufficient to drive the entire pigment cell differentiation gene battery. Network structures of this form are typically wired with coherent feedforward regulatory linkages (Davidson, 2010) and have a tendency to lack direct repression. Instead, the decisions of where and when to deploy a particular gene battery are controlled by upstream regulators, specifically by the cis-regulatory architecture that regulates their expression. The GRN responsible for skeletogenesis is comparatively much deeper. In SM specification, initial spatial inputs, i.e. the double negative regulatory gate, are interpreted by the cis-regulatory architectures controlling *alx1, tbrain, ets1/2* and *tel.* These genes, once expressed, drive a cascade of downstream events that include the expression of circuitry for specification state lockdown, for receiving inductive signals, for mediating EMT transition and subsequent migration, and for driving the expression of the skeletogenic differentiation battery. By turning on *gcm* in the skeletogenic precursors at the point of initial specification, we have introduced a differentiation program in these cells at the same time as they are just beginning to become specified as skeleton. By linking it to the cis-regulatory architecture controlling *tbr*, we have freed *gcm* from the regulatory wiring designed to exclude its expression from the SM lineage. And, as would be expected from a simple differentiation gene battery structure, this has allowed the unrestricted deployment of the pigment-cell differentiation battery. In support of this we observe that *pks* expression occurs even in gcm-expressing clones that continue to participate in syncytium and develop into skeleton.

Prevalence of exclusion functions in development

A common device used in developmental gene regulatory networks is exclusion of alternate regulatory states by cross-repression. These forms of cross-repression of alternate states occur at or just after specification of the primary state and designed to forbid a terminal differentiation program that shares features of the primary specification state. Here we see that *gcm* expression is capable of repressing skeletogenesis in a very similar mesodermal context to that in which the endogenous gene expresses. Consequently we find several SM-specific genes that are downregulated. Among these, *alx1* and *ets1* are expressed at the earliest point of specification of the skeletogenic lineage and when removed have the most severe effects on skeletogenesis. This repressive linkage alone seems however to be unnecessary for normal development as injection of morpholino against endogenous *gcm* does not lead to an expansion of *alx1* to the NSM lineage. This does not exclude, however, the presence of additional parallel mechanisms that may exist to prevent skeletogenesis in the NSM. Nevertheless, it is clear that *gcm* acts at the top of the SM regulatory network.

Use of recombinant BACs to effect wholesale transfer of gene-regulatory architecture

This work has demonstrated the utility of recombinant BACs as tools for reengineering regulatory networks. The use of BACs has several advantages: i)_they show very low ectopic expression compared to smaller constructs; ii) They are large enough to contain both the gene itself and most, if not all, of its cis-regulatory controls, and are therefore capable of driving synthetic genes in a spatial and temporal manner that closely resembles their native genes; iii) when injected, BAC constructs appear to show a greater tendency for early incorporation, in some cases as soon as the 2-cell state. In the sea urchin, the first cleavage occurs along the animal/vegetal axis, producing the left and right half of the embryo. This has made it possible to take advantage of the mosaicism of incorporation by allowing both perturbed and wild-type regulatory states to run in different halves of the same embryo. And iv), by insertion of fluorescent reporters, the BACs also become powerful tools as in-vivo, real-time detectors of specification state.

MATERIALS AND METHODS

Cloning of Tbrain-BAC GFP reporter and Tbrain-BAC GCM expression construct.

A 138 kb BAC clone Sp_031J08_L was identified from a *Strongylocentrotus purpuratus* genomic DNA library from the Sea Urchin Genome Resource (Andy Cameron http://sugp.caltech.edu). This clone contains the entire Tbrain coding sequence, and the start-site of transcription is flanked by at least 60 kb genomic sequence on each side. A full-length spGCM cDNA clone 4I5 was isolated from a 15 hr *S. purpuratus* cDNA library. It contains the complete spGCM coding sequence and 3'UTR. Recombinant BAC cloning was used to generate Tbrain-GCM/GFP expression constructs as described (Lee et al., 2001). GFP or spGCM coding sequence was cloned into a vector upstream of a kanamycin-resistance gene flanked by flp-recombinase sites. Recombination target sequences, roughly 150 bp in length on each end, were cloned upstream of the GCM or GFP coding sequence and downstream of the kanamycin resistance marker. Homologous recombination was used to replace 150 bp of Tbrain sequence (40 bp of 5'UTR and 110 bp of coding sequence) with the coding sequence and 3'UTR of SpGcm. This construct was called the Tbrain::Gcm BAC. A Tbrain:GFP BAC was similarly constructed using GFP coding sequence containing an SV40 3'UTR poly-adenylation tail and was coinjected with Tbrain::Gcm BAC to act as a reporter of exogenous Gcm expression.

The upstream recombination target sequence on the Tbrain BAC was: TTTCGGAAAAAGTGTTAAAATCGCAGTGAGAATTTCATCAGCGTTCGCGCCTTCTCGCTTCTGTGTTTATCCATGTAATTTGTGACTGAATTTTCGCACTCCGACTCTAACCCTAATTTAAAGGGATTGAATTCTAACGCCTTCGCGC.

The downstream target sequence on the Tbrain BAC was: TGAAGATGAGAATCTTGATAGAGATGACGGGAGCAATGGATCTGAAGATACCAACTGCGAAAAGTCAACAGTCGAACAATTTCACACCAATAAATTAATTTCAAACGCTGATCATAACGTCGGGGATCCAAATAACGACTACCCTTGC

Whole-mount in-situ hybridization

Single and 2-color WMISH were performed as described (Minokawa et al., 2005) for detection of DIG- and DNP-labeled probes. In two-color WMISH, DNP-labeled probes were detected with NBT/BCIP staining solution. The reaction was stopped by washing with MOPS buffer, and alkaline phosphatase activity of the anti-DIG-AP Fab fragments was inactivated by glycine-HCl treatment. A second stain was performed on DIG-labeled probes using INT/BCIP staining solution. Embryos were finally transferred to 70% glycerol and imaged.

Microinjections

Tbrain-GCM and Tbrain-GFP BAC constructs: BAC DNA constructs were linearized using the homing endonuclease, PI-SceI to produce 140 kb fragments. A single PI-SceI exists on the pBace3.6 vector used to construct the purpuratus genomic BAC DNA library. BAC DNA were injected into fertilized eggs as described previously (Lee et al., 2007) except that no carrier genomic DNA was used. Injection volumes were in the range of 5-10 picoliters and concentration at 500 copies per pl.

Pks-GFP construct: a reporter construct driving pigment-cell-specific expression of dsRed was cloned. The reporter contains 2.0 kb of genomic DNA upstream of the start site of SpPks transcription and includes the basal promoter and a proximal cis-regulatory module that is capable of recapitulating endogenous Pks expression at 24 and 48 hours post fertilization (Calestani and Rogers, 2010). The primers used to amplify the pks genomic DNA were: upstream primer: 5'-TCCCTCTTTCTCTCCCACTCTC-3',

downstream primer: 5'-CTCTGTTTCTTGCTACAACTCTC-3'

Short Tbrain-GCM/GFP constructs: 5 kb Tbrain constructs containing all the cis-regulatory information necessary to recapitulate embryonic Tbrain expression were PCR-amplified from the BAC-GFP and BAC-GCM constructs (left primer: 5'-TCGGAACGATACGAAAACTTTG-3' right primer: 5'-ACTGCCTCCCTGTTTGAGAA-3').

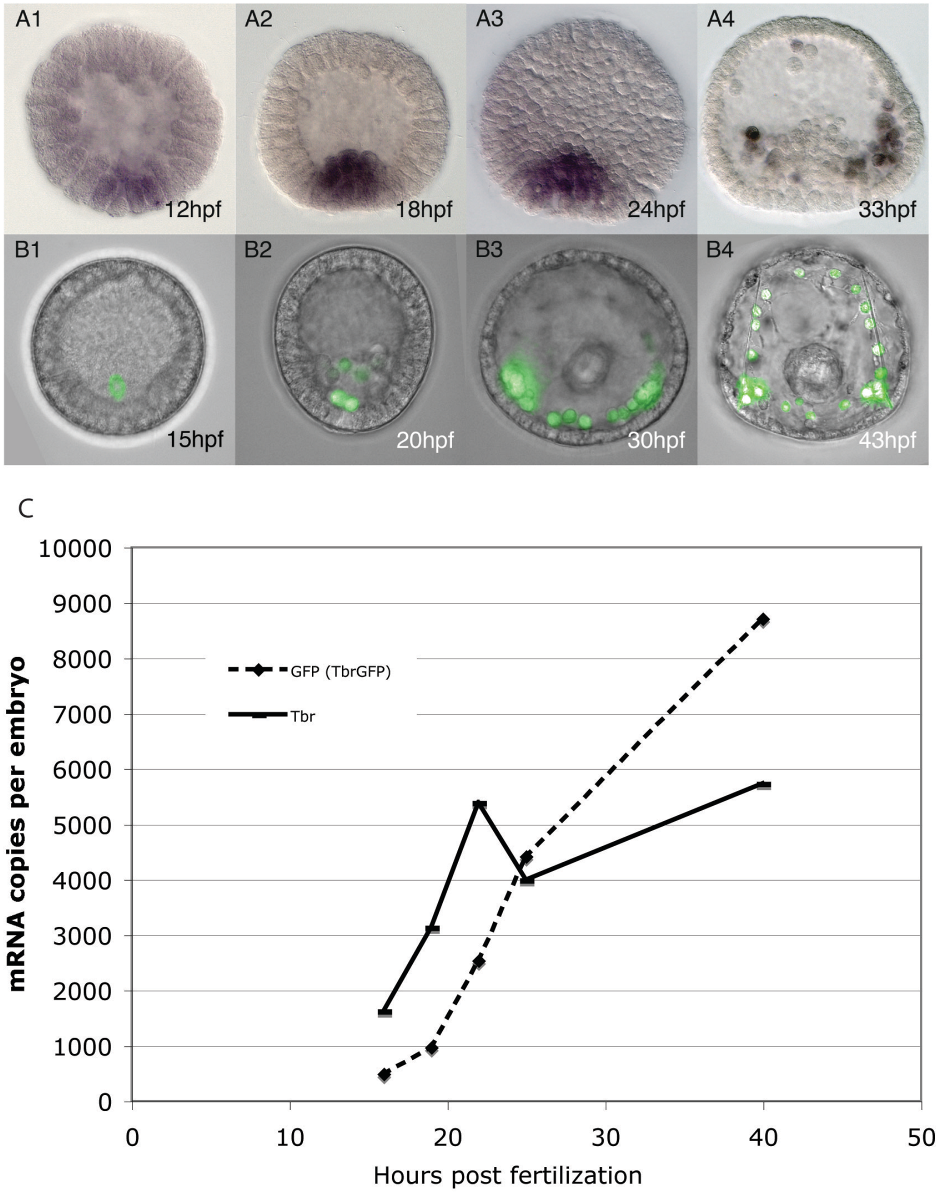
Small constructs were injected as described (Lee et al., 2007). Injection volumes were in the range of 5-10 picoliters and at a concentration of 1000 copies per pl.

Injected construct	Embryos	Time	Activity	Skeletogenic lineage only	Blastocoelar	Pigment/aboral ectoderm
Tbrain-GCM Tbrain-GFP BAC Plate 1-6 e5-63	74	48hpf	33/74	16 [48%]	1 (tip of archenteron) [3%] 11 (blasto) [33%]	5 (abo-ecto + pigment) [15%]
Plate1 e6-145 (F1)	80	48hpf	23/80	9 [39%]	6 (PMC + blasto) [26%] 2 blasto only [9%]	5 (pigment cell + pmc) [22%] 1 pigment only [4%]
Plate2 e6-145 (F2)	62	48hpf	30/62	12 [40%]	7 (pmc + blasto) [23%] 4 blasto only [13%]	6 (pigment cell + pmc) [20%] 1 pigment only [3%]
Tbrain-GFP BAC only Plate 7-12 e5-63	45	48hpf	39/45	31 [79%]	8 [21%]	0 [0%]
Plate 5 e6-145 (F1)	84	48hpf	58/84	54 [93%]	4 [7%]	0 [0%]
Plate 6 e6-145 (F2)	81	48hpf	28/81	22 [79%]	3 [11%]	3 [11%]

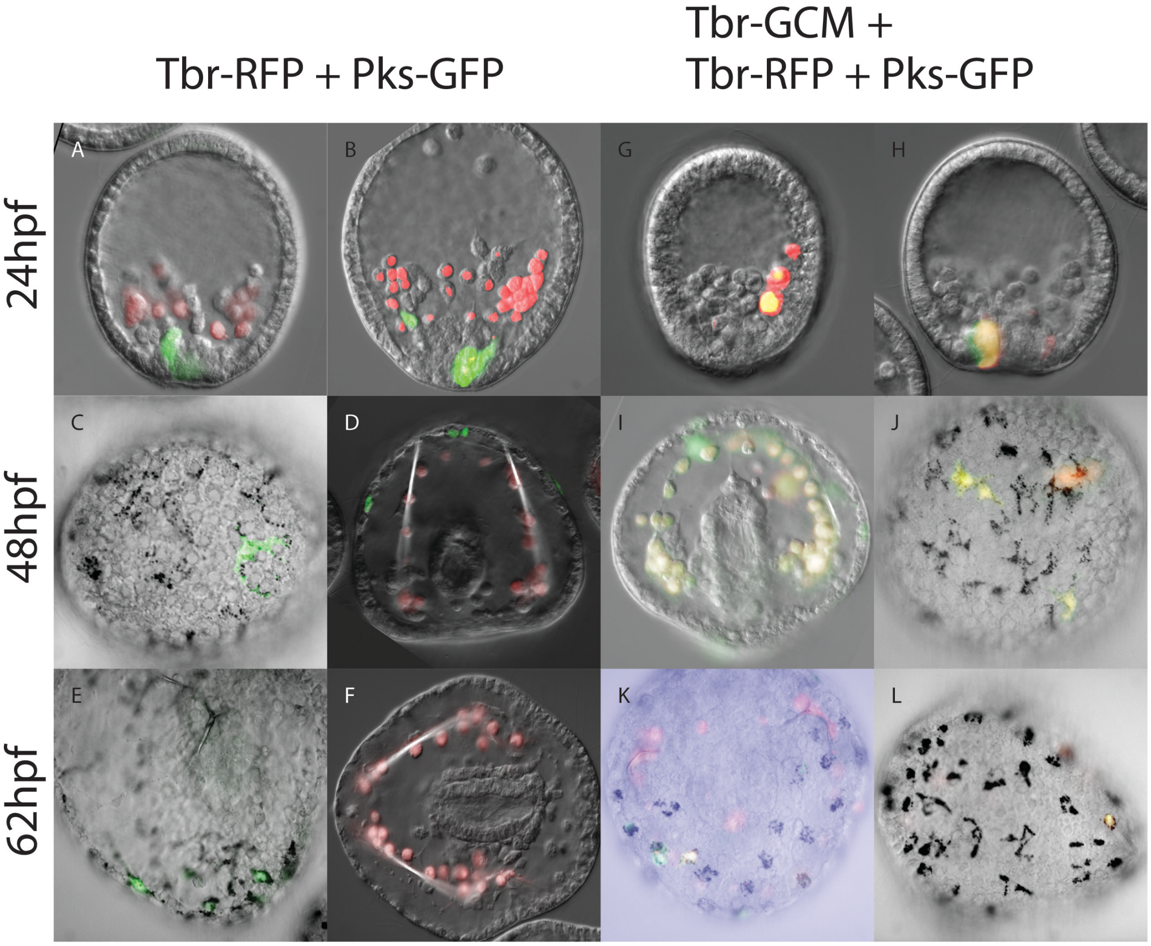
Table 1.1. Expression of Tbrain::GFP and Tbrain::GCM constructs at mid/late gastrula stage (48 hpf). Three categories of GFP expression patterns were scored. Skeletogenic cell expression includes complete expression in ring of fused skeletal mesenchyme. Blastocoel expression includes any morphologically round fluorescent cells seen in the blastocoel. Pigment/aboral ectoderm includes any cells expressing in the aboral ectoderm (some of which express pigment), as well as any unpigmented cells just below the surface of the aboral ectoderm.

		female	RFP+GFP- syncytium GFP+RFP+ aboral ectoderm	RFP+GFP- Syncytium GFP+RFP + blastocoelar	GFP+RFP+ syncytium	GFP+RFP+ Blastocoelar cells	GFP+RFP+ Aboral ectoderm	RFP+GFP- syncytium	RFP+GFP- ectopic blastocoelar
Plate4:70 TbrGCM TbrRFP PksGFP	48	1	7 pigmented 8 unpigmented	7	3	1	5 pigmented 0 unpigmented	8	3
Plate3:91: TbrGCM TbrRFP PksGFP	48	2	6 pigmented 1 unpigmented	6	6	3	5 pigmented 4 unpigmented	7	1
Plate6:84: TbrRFP PksGFP	48	2	0	0	0	0	0	54	4

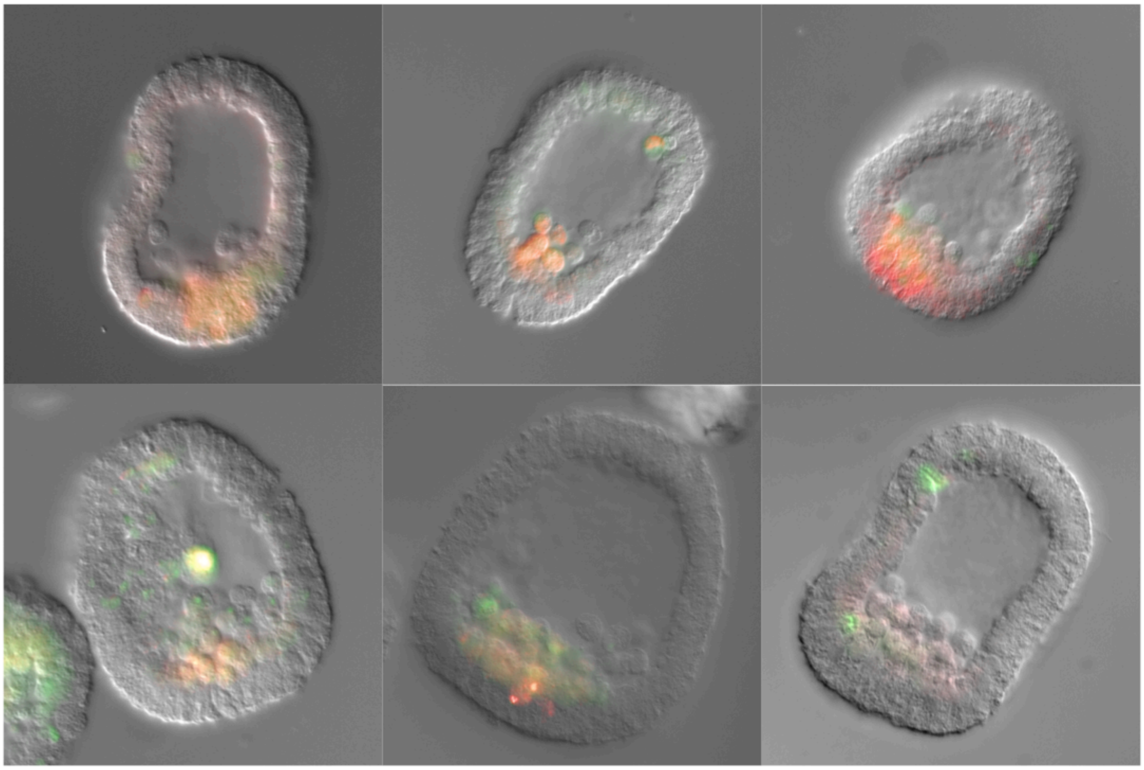
Table1.2. Expression of Tbrain::RFP and Pks::GFP constructs in the presence or absence of coinjected Tbrain::GCM BAC at mid/late gastrulate stage



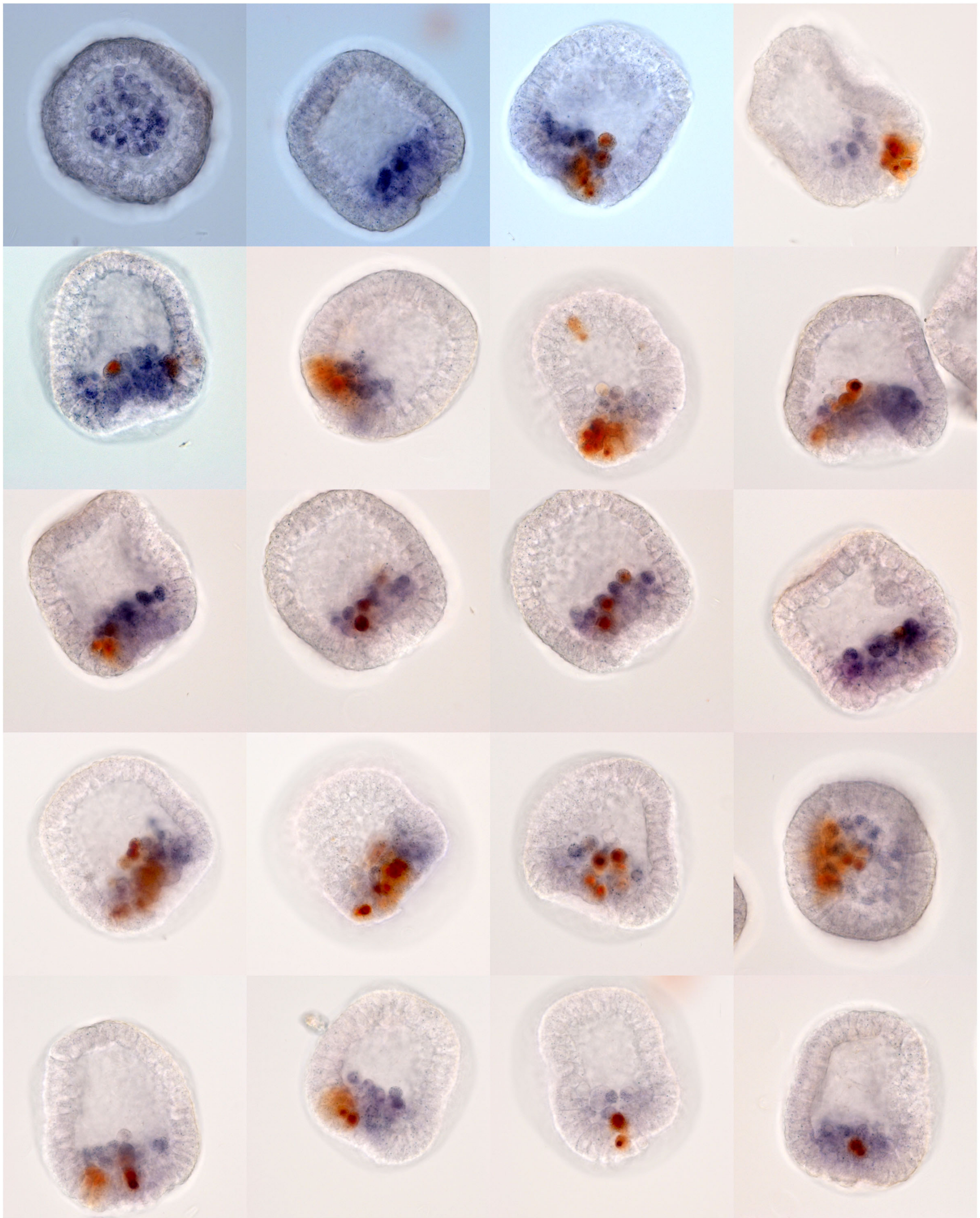
Supplemental Figure 1.1. Expression of GFP reporter driven by Tbrain BAC. (A1-A4) WMISH of uninjected embryos, showing staining for endogenous expression of Tbrain. (B1-B4) GFP fluorescence in embryos injected with Tbrain::GFP BAC reporter construct at 15, 20, 30 and 43 hours post-fertilization. GFP expression at 15 hpf shows mosaic incorporation characteristic of DNA reporter injections. At 30 hpf, primary mesenchyme cells form a syncytium, leading to the even distribution of GFP across all fused cells. (C) Endogenous Tbrain and GFP transcript abundance, measured by QPCR. The copy number of transcripts was calculated by comparing against simultaneously measured ubiquitin mRNA. Ubiquitin has an abundance of roughly 300000 copies per embryo.



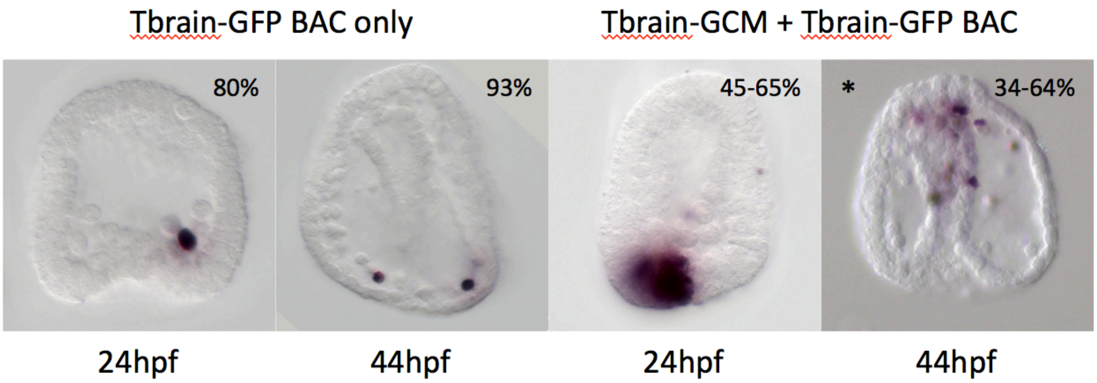
Supplemental Figure 1.2. Tbrain-RFP and Pks-GFP small construct reporters were cloned such that their expression matches that of endogenous Tbrain and PKS. Pks is a differentiation gene that is involved in pigment synthesis and is expressed in the subset of SMC cells that eventually migrate to the aboral ectoderm. Hence these constructs act as markers for the PMC and pigment cell types. These two were coinjected either in the presence or absence of a Tbrain-GCM small construct and observed during development. In (A-F) embryos were coinjected with only Tbr-RFP and Pks-GFP. (A,C,E) show expression of Pks-GFP reporter in SMCs and eventually in pigment cells in the aboral ectoderm. (B,D,F) shows expression of Tbrain-RFP construct in ingressing PMCs and eventually in the skeleton. In (G-L) embryos were coinjected with the Tbrain-GCM as well as Tbrain-RFP and Pks-GFP. G,I, show coexpression of RFP and GFP in PMCs and skeletogenic cells and (H,J,K,L) show coexpression of RFP and GFP in SMCs and pigment cells.



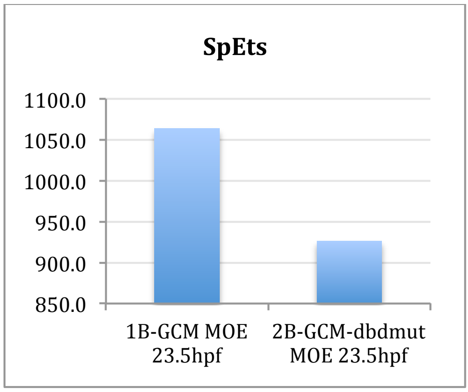
Supplemental figure 1.2: *tbr* and *alx1* BAC reporter expression in the presence of synthetic *gcm* expression at 20-24hpf (near mesenchyme blastula stage). *alx1* BAC reporter, red; *tbr* BAC reporter, green. Top row) injection of *tbr* and *alx1* BAC reporters only. Detectors are coexpressed in large micromeres. 42 out of 42 embryos showed strong expression of both detectors exclusively in the large micromeres Bottom row) coinjection of *tbr* and *alx1* detectors with Tbrain:Gcm BAC. Expression of BAC construct leads to downregulation of both detectors. 0 of 45 embryos showed strong expression of both reporters. Expression patterns were weak and imaging required longer exposure. 2 of 45 embryos showed weak *alx* reporter expression in large micromeres and 3 of 45 embryos showed weak *tbr* reporter in large micromeres.

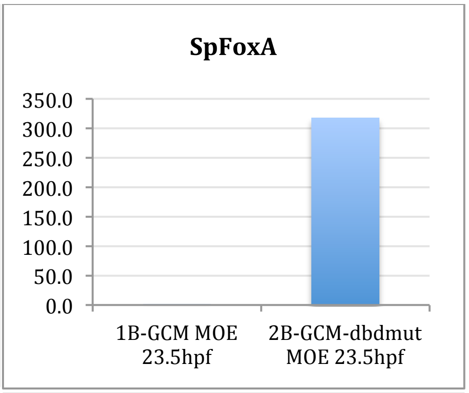


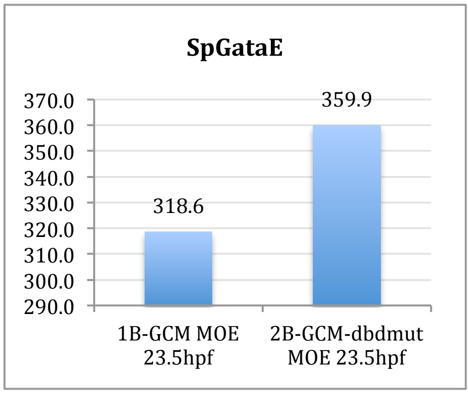
Supplemental figure 1.3: Alx expression in Tbrain:GCM BAC-injected embryos. Alx-dig probe, purple; synthetic gcm probe, orange

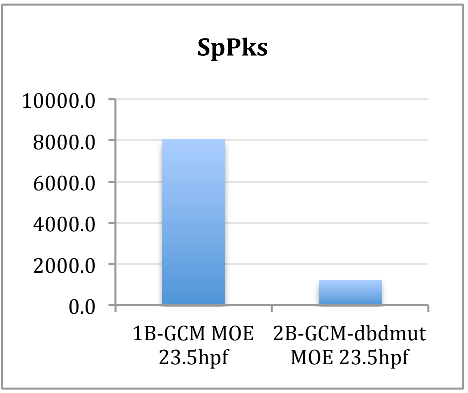


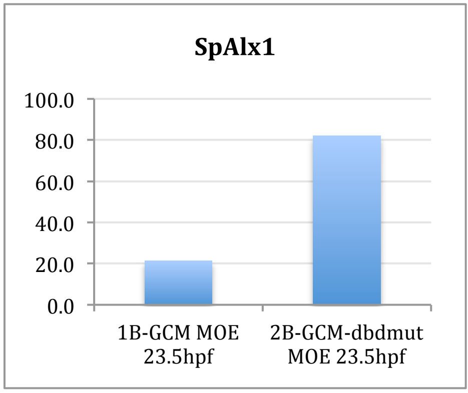
Supplemental Figure 1.4. Whole mount in-situ of Tbrain-GCM BAC injected embryos at mesenchyme blastula stage and late gastrula stage

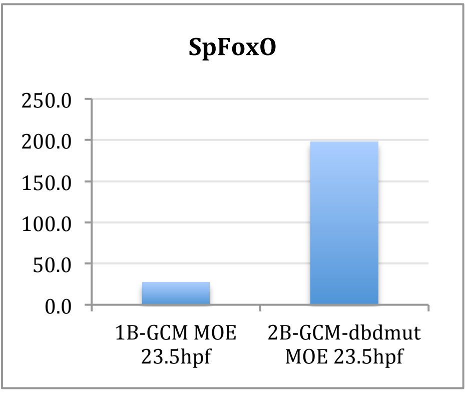

SpEts
1100.0
1050.0
1000.0
950.0
900.0
850.0
1B-GCM MOE 23.5hpf
2B-GCM-dbdmut MOE 23.5hpf

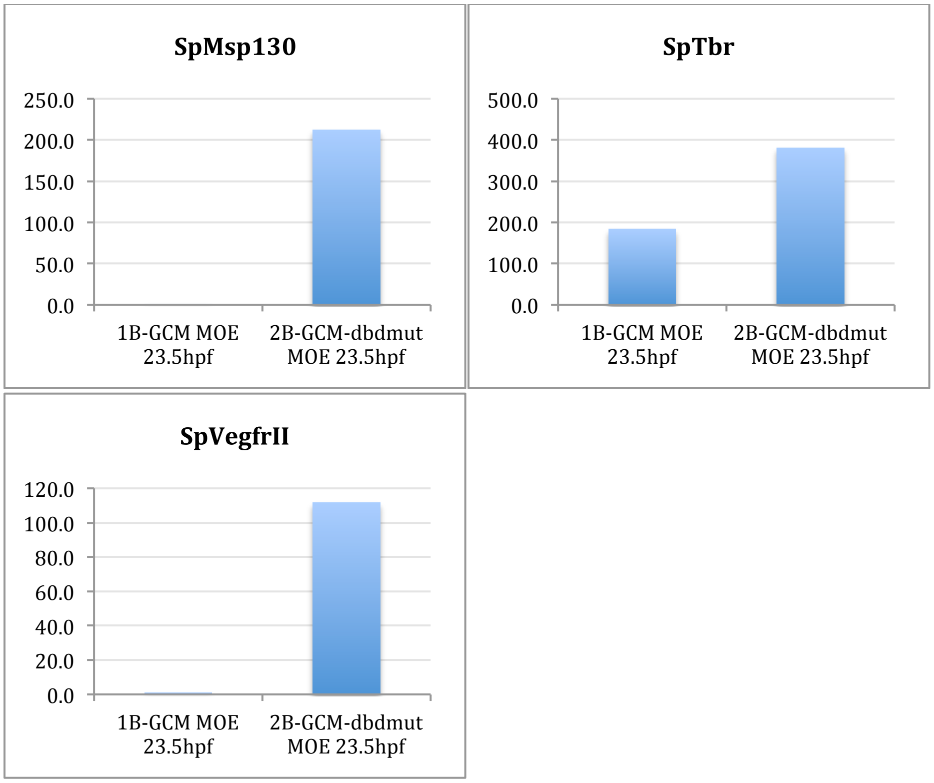

SpFoxA
350.0
300.0
250.0
200.0
150.0
100.0
50.0
0.0
1B-GCM MOE 23.5hpf
2B-GCM-dbdmut MOE 23.5hpf


SpGataE
370.0
360.0
350.0
340.0
330.0
320.0
310.0
300.0
290.0
318.6
359.9
1B-GCM MOE 23.5hpf
2B-GCM-dbdmut MOE 23.5hpf


SpPks
10000.0
8000.0
6000.0
4000.0
2000.0
0.0
1B-GCM MOE 23.5hpf
2B-GCM-dbdmut MOE 23.5hpf


SpAlx1
100.0
80.0
60.0
40.0
20.0
0.0
1B-GCM MOE 23.5hpf
2B-GCM-dbdmut MOE 23.5hpf


SpFoxO
250.0
200.0
150.0
100.0
50.0
0.0
1B-GCM MOE 23.5hpf
2B-GCM-dbdmut MOE 23.5hpf



Supplemental Figure 1.5. Effect of gcm mRNA overexpression at mesenchyme blastula stage. Embryos were injected with mRNA of wild-type *gcm* or a *gcm* with a mutated DNA-binding domain. Expression reported as copies mRNA per embryo. *Gcm* DNA-binding domain mutant contains the mutation N65D. mRNA were injected at roughly 400000 copies per embryo.

REFERENCES

Altmann, C.R., Chow, R.L., Lang, R.A., Hemmati-Brivanlou, A., 1997. Lens induction by Pax-6 in Xenopus laevis. Developmental Biology 185, 119-123.

Calestani, C., Rogers, D.J., 2010. Cis-regulatory analysis of the sea urchin pigment cell gene polyketide synthase. Developmental Biology 340, 249-255.

Croce, J., Lhomond, G., Lozano, J.C., Gache, C., 2001. ske-T, a T-box gene expressed in the skeletogenic mesenchyme lineage of the sea urchin embryo. Mechanisms of Development 107, 159-162.

Davidson, E.H., Erwin, D.H., 2006. Gene regulatory networks and the evolution of animal body plans. Science 311, 796-800.

Davidson, E.H., Rast, J.P., Oliveri, P., Ransick, A., Calestani, C., Yuh, C.-H., Minokawa, T., Amore, G., Hinman, V., Arenas-Mena, C., Otim, O., Brown, C.T., Livi, C.B., Lee, P.Y., Revilla, R., Schilstra, M.J., Clarke, P.J.C., Rust, A.G., Pan, Z., Arnone, M.I., Rowen, L., Cameron, R.A., McClay, D.R., Hood, L., Bolouri, H., 2002. A provisional regulatory gene network for specification of endomesoderm in the sea urchin embryo. Developmental Biology 246, 162-190.

Elowitz, M.B., Leibler, S., 2000. A synthetic oscillatory network of transcriptional regulators. Nature 403, 335-338.

Ettensohn, C.A., Illies, M.R., Oliveri, P., 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.

Gao, F., Davidson, E.H., 2008. Transfer of a large gene regulatory apparatus to a new developmental address in echinoid evolution. Proc Natl Acad Sci USA 105, 6091-6096.

Halder, G., Callaerts, P., Gehring, W.J., 1995. Induction of ectopic eyes by targeted expression of the eyeless gene in Drosophila. Science 267, 1788-1792.

Howard-Ashby, M., Materna, S.C., Brown, C.T., Chen, L., Cameron, R.A., Davidson, E.H., 2006. Gene families encoding transcription factors expressed in early development of Strongylocentrotus purpuratus. Developmental Biology 300, 90-107.

Lee, E.C., Yu, D., Martinez de Velasco, J., Tessarollo, L., Swing, D.A., Court, D.L., Jenkins, N.A., Copeland, N.G., 2001. A highly efficient Escherichia coli-based chromosome engineering system adapted for recombinogenic targeting and subcloning of BAC DNA. Genomics 73, 56-65.

Lee, P.Y., Nam, J., Davidson, E.H., 2007. Exclusive developmental functions of gatae cis-regulatory modules in the Strongylocentrorus purpuratus embryo. Developmental Biology 307, 434-445.

Materna, S.C., Howard-Ashby, M., Gray, R.F., Davidson, E.H., 2006. The C2H2 zinc finger genes of Strongylocentrotus purpuratus and their expression in embryonic development. Developmental Biology 300, 108-120.

McMahon, A.P., Flytzanis, C.N., Hough-Evans, B.R., Katula, K.S., Britten, R.J., Davidson, E.H., 1985. Introduction of cloned DNA into sea urchin egg cytoplasm: replication and persistence during embryogenesis. Developmental Biology 108, 420-430.

Minokawa, T., Wikramanayake, A.H., Davidson, E.H., 2005. cis-Regulatory inputs of the wnt8 gene in the sea urchin endomesoderm network. Developmental Biology 288, 545-558.

Oliveri, P., Tu, Q., Davidson, E.H., 2008. Global regulatory logic for specification of an embryonic cell lineage. Proc Natl Acad Sci USA 105, 5955-5962.

Peter, I.S., Davidson, E.H., 2009. Modularity and design principles in the sea urchin embryo gene regulatory network. FEBS Lett 583, 3948-3958.

Ransick, A., Rast, J.P., Minokawa, T., Calestani, C., Davidson, E.H., 2002. New early zygotic regulators expressed in endomesoderm of sea urchin embryos discovered by differential array hybridization. Developmental Biology 246, 132-147.

Revilla-i-Domingo, R., Minokawa, T., Davidson, E.H., 2004. R11: a cis-regulatory node of the sea urchin embryo gene network that controls early expression of SpDelta in micromeres. Developmental Biology 274, 438-451.

Rizzo, F., Fernandez-Serra, M., Squarzoni, P., Archimandritis, A., Arnone, M.I., 2006. Identification and developmental expression of the ets gene family in the sea urchin (Strongylocentrotus purpuratus). Developmental Biology 300, 35-48.

Sweet, H.C., Gehring, M., 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, 1945-1955.

Sweet, H.C., Hodor, P.G., Ettensohn, C.A., 1999. The role of micromere signaling in Notch activation and mesoderm specification during sea urchin embryogenesis. Development 126, 5255-5265.

Wahl, M., Hahn, J., Gora, K., Davidson, E., 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. Developmental Biology. 2009 Nov 15;335(2):428-41

CHAPTER 2

Precise *Cis*-Regulatory Control of Spatial and Temporal Expression of the *alx-1* Gene in the Skeletogenic Lineage of *S. purpuratus*

Sagar Damle and Eric H. Davidson

(In preparation, *Developmental Biology*)

ABSTRACT

Deployment of the gene regulatory network (GRN) responsible for skeletogenesis in the embryo of the sea urchin *Strongylocentrotus purpuratus* is restricted to the large micromere lineage by a double negative regulatory gate. The gate consists of a GRN subcircuit composed of the *pmar1* and *hesC* genes, which encode repressors and are wired in tandem, plus a set of target regulatory genes under *hesC* control. The skeletogenic cell state is specified initially by micromere-specific expression of these regulatory genes, *viz*. *alx1, ets1, tbrain,* and *tel*, plus the gene encoding the Notch ligand Delta. Here we use a recently developed high-throughput methodology for experimental *cis*-regulatory analysis to elucidate the genomic regulatory system controlling *alx1* expression in time and embryonic space. The results entirely confirm the double-negative gate control system at the *cis*-regulatory level, including definition of the functional HesC target sites, and add the crucial new information that the drivers of *alx1* expression are initially Ets1, and then Alx1 itself plus Ets1. *Cis*-regulatory analysis demonstrates that these inputs quantitatively account for the magnitude of *alx1* expression. Furthermore,

the Alx1 gene product not only performs an auto-regulatory role, promoting a fast rise in *alx1* expression, but also, when at high levels, it behaves as an autorepressor. A synthetic experiment indicates that this behavior is probably due to dimerization. In summary, the results we report provide the sequence-level basis for control of *alx1* spatial expression by the double negative gate GRN architecture, and explain the rising, then falling, temporal expression profile of the *alx1* gene in terms of its auto-regulatory genetic wiring.

INTRODUCTION

Developmental gene regulatory networks (GRNs) are models that explain embryonic specification functions in terms of a hierarchical matrix of genomically encoded information processing events. The GRN that encodes pre-gastrular development of the *S. purpuratus* large-micromere/skeletogenic mesenchyme (SM) lineage is to date among the most complete and well studied (Davidson et al., 2002; Oliveri et al., 2002; Oliveri et al., 2003; Oliveri et al., 2008). The specification of the large micromeres is initiated by action of a double-negative regulatory gate, whereby micromere specific expression of the repressor gene *pmar1* in turn represses transcription of the ubiquitously-driven repressor gene, *hesc* (Oliveri et al., 2003; Revilla-i-Domingo et al., 2007; Oliveri et al., 2008). This regulatory gate can be shown to operate as a logic processing device (Peter and Davidson, 2009). It is also of evolutionary importance, as it has been considered the focal point in the GRN for the redeployment of the pre-existing adult skeletogenic apparatus to the micromere lineage early in the evolutionary divergence of the euchinoids (Gao and Davidson, 2008). Understanding the sequence basis of the mechanism by which HesC repression unlocks the skeletogenic program will illuminate the pathway by which such network co-options may have occurred.

Alx1, *ets1, tbrain,* and *tel* are the earliest transcription factors defining the definitive zygotic skeletogenic micromere (SM) regulatory state, and together with the gene encoding the Notch ligand Delta, these genes are expressed immediately downstream of the double negative regulatory gate. In previous work the sequence basis of HesC repression, and thus in the SM lineage the release from this repression, has been identified at the genomic *cis*-regulatory level for the *tbrain* gene (Wahl et al, 2008) and the *delta* gene (Revilla et al, 2007; Smith et al, 2008). But despite its key importance for the subsequent developmental processes of the SM lineage (Ettensohn et al., 2003), no *cis*-regulatory information has been available for the *alx1* gene. *Alx1* encodes the first invertebrate member of the Cart1/Alx3/Alx4 family of Paired-class homeodomain proteins, also known as Group-I Aristaless-like factors (Ettensohn et al., 2003). In addition to its homeodomain, the protein encoded by the *Strongylocentrotus alx1* gene shares with vertebrate CART family members the presence of a charged domain near the N-terminus, an OAR/Aristaless domain at the C-terminus, and a generally proline-rich primary sequence. Alx transcription factors appear to share an ancient, conserved role in skeletogenic development. The *alx1* gene is expressed in both juvenile sea urchin and sea star skeletonization centers (Gao and Davidson, 2008), as well as in the sea urchin embryo, while several Group-I *aristalless* like genes are expressed during vertebrate embryogenesis in the mesenchymal cells that form the craniofacial and appendicular skeleton (Beverdam and Meijlink, 2001; Qu et al., 1997). Loss-of-function mutations in these genes lead to defects in skeletal elements in mice. Though the downstream effector molecules for skeletogenesis in echinoderms and vertebrates are different (the biominerals per se are non-homologous), these similar expression patterns may reflect functional conservation of regulatory cassettes controlling skeletogenic state specification in the ancestral deuterostome.

Alx1 is first expressed in the large daughters of 4th cleavage micromeres and its spatial expression is restricted to the descendants of this cell lineage, the skeletal mesenchyme (SM), for the remainder of embryogenesis. The quantitative temporal profile of the *alx1* expression pattern is fairly complex, as was first observed by P. Oliveri (unpublished), and illustrated here in Fig 2.1. Expression begins around 7.5 hours post fertilization (hpf), and peaks twice during embryogenesis, at first sharply at pre-hatching blastula stage (10-12 hpf) and then more gradually at mesenchyme blastula stage (23-25 hpf). Previous studies based on morpholino antisense interference have suggested that *alx1* is driven in SM cells by the Ets1 transcription factor (Ettensohn et al., 2003; Sharma and Ettensohn, 2010; Oliveri et al, 2008).

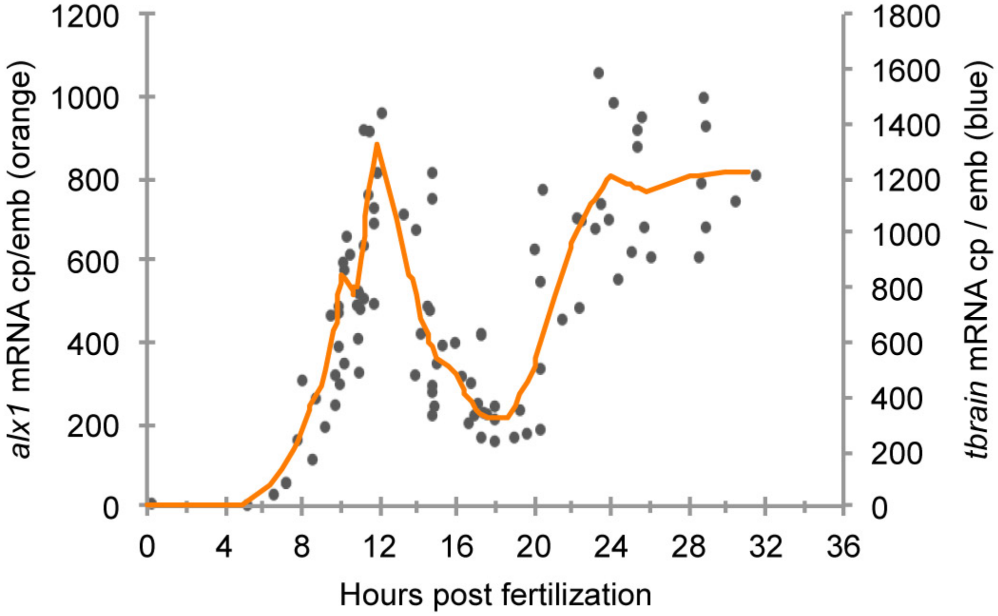


Figure 2.1. High-density timecourse of endogenous *alx1* expression. Measurements of *alx1* mRNA abundance were compiled from multiple (n=8) experiments over the course of the first 30 hours of development and smoothed by LOWESS regression (orange line).

GRN models can ultimately be validated by *cis*-regulatory analysis, in which the predicted target sites are identified and their predicted functionalities demonstrated. This level of structure/function analysis immediately identifies the genomic regulatory code, the functional meaning of which is explicitly predicted in the GRN model, and *cis*-regulatory analysis is also the final arbiter of direct vs. indirect genetic interactions. Here we deconstruct the *alx1* expression pattern during early development into activation and repression components by identifying the genomic regulatory sequences responsible for these functions. Thus we have experimentally identified, and by mutation functionally characterized the genomic target sites responsible for direct spatial repression by the *hesc* gene product, and for activation by Ets1. In addition, we demonstrate that Alx1 protein is both an immediate, direct, auto-activator, and at higher concentrations an auto-repressor, and reveal the biochemical and gene-regulatory network architectural features that permit these opposing roles. These regulatory interactions provide an explanation for both the lineage-specific spatial expression of the *alx1* gene and its kinetic expression profile.

MATERIALS AND METHODS

Injection and scoring of reporter constructs

Sea urchin eggs and sperm were isolated and prepared for injection as described (Cheers and Ettensohn, 2004). Small constructs were injected in a solution containing 120mM KCl, and 30ng/ul of carrier DNA. BAC-GFP constructs were injected without carrier DNA. In barcoded GFP reporter experiments, multiple DNA constructs were mixed and co-injected at a total concentration of 0.9ng/ul (roughly 110 total copies per 2pl), and injection volume per egg was approximately 10-20pl. mRNA constructs were injected without carrier DNA at concentrations ranging from 1ug/ul to 100ng/ul.

Isolation of Alx1 BAC and phylogenetic footprinting

An *alx1* BAC (Sp_BAC_042I08_L) was isolated from a *Strongylocentrotus purpuratus* BAC library as described (Lee et al., 2007), using a partial cDNA probe. The BAC was mapped to get an estimate of the minimum distance between the *alx1* coding sequence and the termini of the insert. Mapping was performed by digesting the BAC with Kpn1 and gel-purifying the individual restriction fragments. These fragments were used as templates for QPCR. Each fragment was assayed for the presence of vector sequence (pBACe 3.6) and for *alx1* exon sequences. The mapping step was used to preclude BACs that are not desirable for *cis*-regulatory analysis because individual restriction fragments contain both vector sequence and *alx1* coding sequence, i.e., in which the *alx1* gene borders the edge of the BAC insert.. Similar procedures were also used to isolate a *Lytechinus variegatus alx1* BAC (Lv_BAC_007J11_L).

Phylogenetic footprinting between the *S.p.* and *L.v. alx1* BAC sequences was performed using SeqComp, and visualized by the Family Relations software package (Brown et al., 2002). Seqcomp was performed using a 50bp window and 80% sequence similarity.

Generation of BAC GFP reporter and deletion constructs by homologous in vitro *recombination.*

The parental BAC is here referred to as *alx1:*GFP BAC. It was generated by homologous recombination as described (Court et al., 2002). The targeting cassette contained the GFP coding sequence, an SV40 poly-adenylation site, and the kanamycin gene, flanked by flp-recombinase target sites. The targeting cassette was amplified using primers with 5' tails homologous to the insertion site as follows (*alx1*-specific targeting sequence underlined):

Alx-GFP-cassette_Forward:

GCCTTTTCTTAGGATTTTGTCGTGCCGAGACTTTACTCAATATTGATGAGCAA
GGGCGAGGAACT

Alx-GFP-cassette_Reverse:

AGTTTACTTACACGTCGCTAAGCACGGCATTGAGGGGTAAAACAATCGAAGA
GCTATTCCAGAAGTAGTGA

Homologous recombination with this cassette replaced the first 36 bp of coding sequence from the 3' portion of *alx1* exon1 with the GFP cassette. After recombination, the kanamycin resistance gene and bacterial regulatory DNA were removed by induction of the *flippase* gene, leaving a 126 bp artifact containing one 45 bp flp-recombinase site downstream of the SV40 3'UTR.

The presence of an extraneous flp site acts as an anchor point for subsequent homologous recombination experiments using the flp-recombinase. Therefore, an alternative strategy involving Galk positive/negative selection (Warming et al., 2005) was used to generate mutational variants of the *alx1:*GFP BAC. A targeting cassette containing galK was amplified using tailed primers containing sequences flanking two putative HesC binding sites as follows (*alx1*-specific targeting sequence underlined):

HesC flanking site Forward:

ACTCTTGACCAATGACCGTGCCCGAAGCCCAGCGGTGTATAATAGCCTGTTGACAATTAATCATC

HesC flanking site Reverse:

GAGCGAGAGTGAAAATCGGCGAGTGCTTCGGCGGAGCGAAGAAACTCAGCACTGTCCTGCTCCTT

Recombinant Galk-containig BACs were screened for proper insertion using primer pairs that bridged the insertion site, and later confirmed by sequencing. A second homologous recombination was performed using a cassette containing the desired mutated sequence and flanked by 150 bp homologous target-sequence. Recombinants were isolated by negative selection for Galk as described.

Generation of cis-regulatory reporter constructs for GFP scoring and QPCR

Cis-regulatory reporter constructs were generated by fusing a putative regulatory sequence to *alx1* basal promoter and to a GFP cassette in two successive fusion PCR steps as described (Hobert, 2002). An adaptamer with the following sequence was used to fuse putative *cis*-regulatory modules to the Alx1 basal promoter:

AGCTTGATATCGAAGTCCTGCAG

The set of 13 "barcoded" GFP vectors that we developed for high throughput *cis*-regulatory analysis (Nam et al., 2010) were individually fused to various regulatory DNA/promoter construct combinations, mixed into the same injection solution, and injected in fertilized eggs. The GFP "barcode " sequence tags are detected independently using specific QPCR primers (Nam et al, 2010). QPCR was used to measure reporter activity of each tag GFP construct quantitatively, and the results were normalized to the number of integrated genomic copies of that tag as described (Revilla-i-Domingo et al., 2004). GFP expression as measured by the abundance of unique tags was then also normalized for minor tag-specific differences in transcript half-life. This was done by assaying the variability of expression of 13 tag reporters when driven by *identical* active cis-regulatory modules. This measurement was repeated 5 times and used as a normalization standard in all tag experiments (Supplemental Figure 2.1). A negative control was constructed by fusing the basal promoter-GFP construct to a series of nonfunctional genomic fragments ~ 500-1000 bp in length. Expression from this construct is used to set a baseline for all expression data (Supplemental Figure 2.2).

Mutation of putative transcription factor binding sites within reporter constructs

Site-specific mutation of reporter constructs was performed using fusion PCR with primers overlapping the target sequence but containing the desired mutation or deletion. Each primer was approximately 45 bp long and included the target site disruption and 20 bp of unmutated flanking sequence. Primer sequences used to generate the following mutants are described in Supplemental Table 2.1: *et1s (x5), hesc proximal, hesc distal, tcf proximal, tcf distal, alx distal (x3), alx-proximal.*

Overexpression of mRNA encoding monomeric and tethered obligate dimer forms of alx1

Monomeric *alx1* mRNA was constructed by amplifying the full-length coding sequence from a population of 11.5 hpf cDNAs. An obligate dimer of *alx1* was generated through fusion PCR by joining two copies of the *alx1* coding sequence with coding sequence for a glycine-serine tether (GGGGS)x3 kindly provided by Joshua Klein, Caltech. Each construct was cloned into p-gemT vector and capped mRNAs were synthesized using the T7 mMessage mMachine RNA Transcription Kit (Ambion) and polyadenylated using the poly-A synthesis kit (Ambion). These synthetic mRNAs were injected into fertilized eggs as described above.

RESULTS

Structure of the alx1 genomic locus

The *alx1* locus can be found on NCBI genomic scaffold NW_001306657. In addition to *alx1*, this scaffold contains three other genes: *alx-related1*, *LOC583266* and *pit54-related* (Figure 2.2a). A comparison of scaffold sequence to *alx1 cDNA* (NM_214644, GLEAN3_25302) revealed that the *alx1* gene contains 6 exons extending over 36kb of genomic DNA: a 251 bp 5'UTR is contained within exon1 and a 3360bp 3'UTR is in exon 6. The Alx1 homeodomain coding region is encoded in exons 2-4. A putative transcriptional start site, containing a canonical TATA box was identified 49bp upstream of the 5'UTR (figure 2.2A).

Recapitulation of endogenous alx1 expression by an alx1 GFP BAC reporter

A 129 kb BAC (042I08_L) containing *alx1* was sequenced and found to include all 6 exons as well as 35 kb of upstream and 57 kb of downstream flanking sequence (Figure 2.2B). Using bacterial homologous recombination, we inserted a GFP reporter cassette within exon1 at the start of the *alx1* coding sequence (Figure 2.2B). When injected into fertilized eggs and assayed for expression during development, the BAC-GFP reporter activity closely resembled endogenous *alx1* expression from early blastula through to late gastrula stage (Figure 2.2 c,d). QPCR of the alx-GFP BAC reporter shows that expression begins at 8 hpf and reaches peak activity of roughly 80 copies GFP per

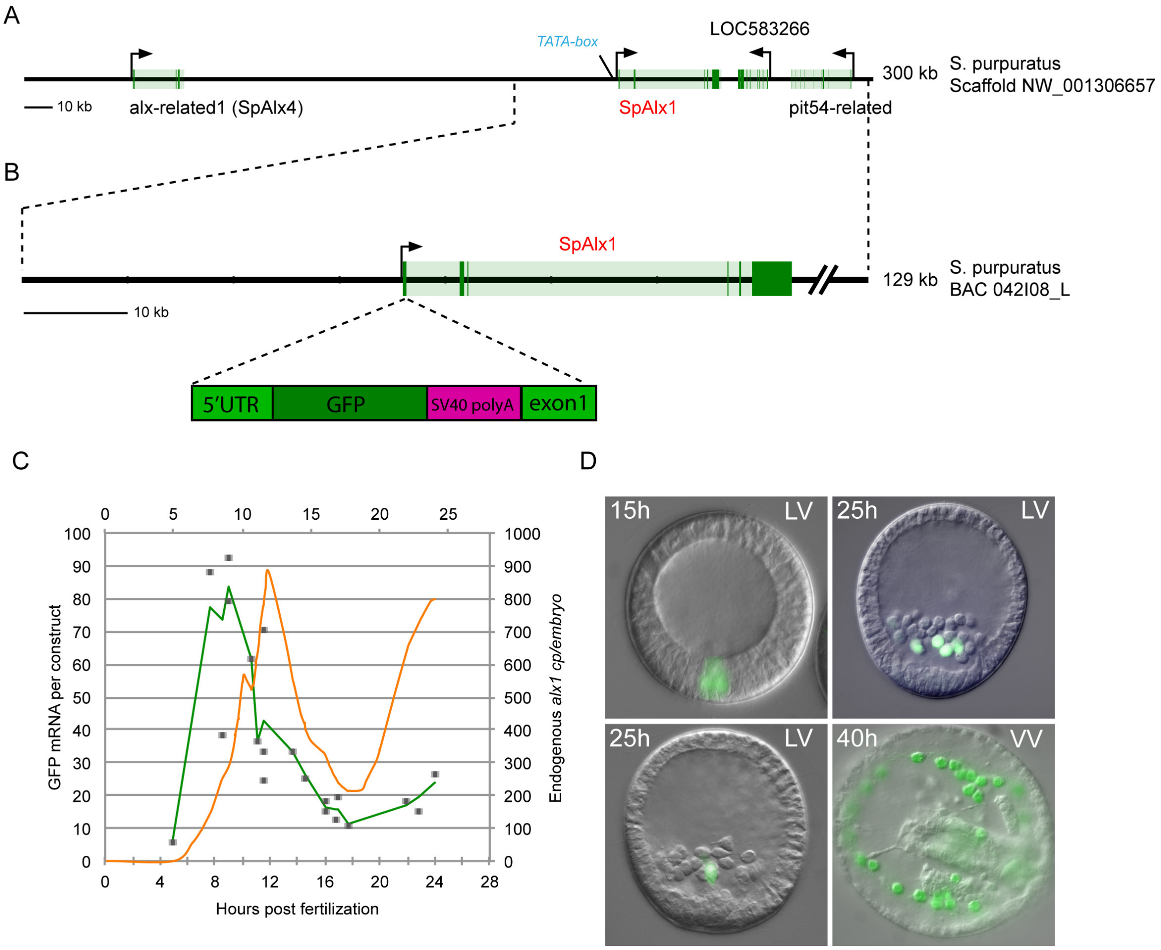
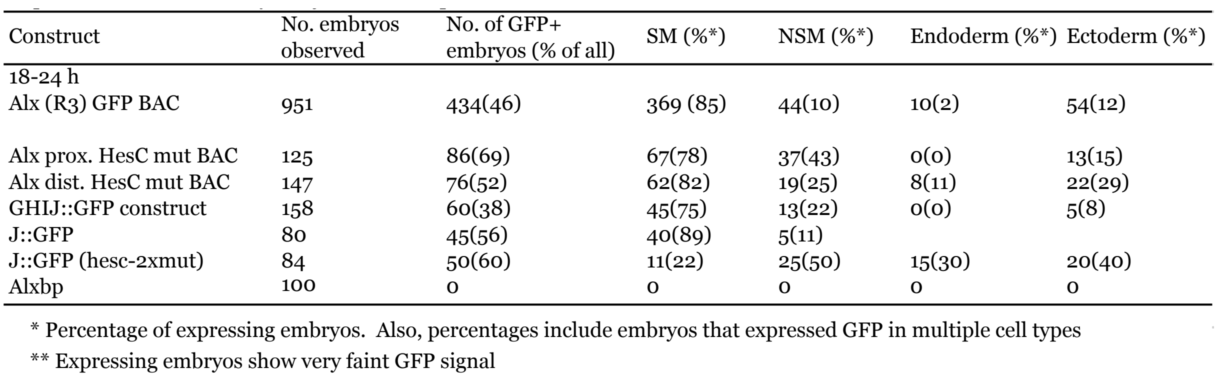


Figure 2.2. Correct spatiotemporal expression of alx-GFP BAC reporter construct. A) Scaffold NW_001306657 contains 4 genes including *alx1*. The alx1-related gene, *sp-alx4* (NCBI Ref Seq. XM_780145, Baylor Gene GLEAN3_22816) lies roughly 160 kb upstream of *alx1* and is oriented in the same direction. Two genes lie within 50 kb downstream of *alx1* and are oriented opposite to *alx1*, LOC583266 and *pit54-related* (NCBI Ref Seq. XM_783163.2). A canonical TATA box sits at -49 bp (blue). B) A 140kb BAC (Sp_042I08_L) isolated from an *S.p.* genomic library was found to contain *alx1*. The coding sequence is flanked upstream by 35 kb of genomic DNA and downstream by 65 kb. A cassette containing GFP coding sequence and SV40polyA 5'UTR was inserted at the start of *alx1* coding sequence as described C) GFP mRNA expression in embryos injected with the Alx-GFP BAC was measured by qPCR from 5-24 hpf. Data was combined from several experiments (n=5) and smoothed using LOWESS regression (green line). Endogenous *Alx1* expression timecourse is overlaid on the secondary axis (orange line). D) BAC-GFP-injected embryos were imaged for GFP fluorescence and overlaid onto DIC pictures at 15, 25 and 40 hpf (top row 15 and 25 hpf, bottom row 25 and 40 hpf). LV, lateral view; VV, ventral view

construct (cp/construct) at 9-10 hpf. Expression drops to 10 cp/construct at 16-18 hpf and is followed by a second smaller peak of 25 cp/construct at 24 hpf (Figure 2.2c). For comparison the kinetics of endogenous *alx1* expression from Figure 2.1 are co-plotted in Fig. 2.2c. The kinetics are slightly different; note that the turnover of *gfp* mRNA is slightly slower than that of *alx1* mRNA, as evident in the falling portions of the respective curves (peak-to-trough 9 h for *gfp* mRNA vs 6 h for *alx1* mRNA), but both display the same sharp rise in mRNA level followed by an abrupt decline. In living injected embryos, GFP fluorescence was initially detected in the large micromeres as early as 12 hpf (several hours are required for GFP protein to fold into its native fluorogenic form in sea urchin egg cytoplasm at 15°). Expression persisted in the large micromeres and their descendants, the SM lineage, throughout the remaining 72 hours of embryonic development (Figure 2.2d), and there was minimal ectopic expression (Table 2.1). These observations demonstrate, that as would be expected from the position of the gene in the BAC, the *alx1* BAC GFP reporter contains all the necessary *cis*-regulatory control apparatus to recapitulate the correct spatial and temporal expression of *alx1* in the embryo.



Construct	No. embryos observed	No. of GFP+ embryos (% of all)	SM (%*)	NSM (%*)	Endoderm (%*)	Ectoderm (%*)
18-24 h						
Alx (R3) GFP BAC	951	434(46)	369 (85)	44(10)	10(2)	54(12)
Alx prox. HesC mut BAC	125	86(69)	67(78)	37(43)	0(0)	13(15)
Alx dist. HesC mut BAC	147	76(52)	62(82)	19(25)	8(11)	22(29)
GHIJ::GFP construct	158	60(38)	45(75)	13(22)	0(0)	5(8)
J::GFP	80	45(56)	40(89)	5(11)		
J::GFP (hesc-2xmut)	84	50(60)	11(22)	25(50)	15(30)	20(40)
Alxbp	100	0	0	0	0	0

* Percentage of expressing embryos. Also, percentages include embryos that expressed GFP in multiple cell types

** Expressing embryos show very faint GFP signal

Table 2.1. Expression of GFP in embryos injected with reporter constructs

Application of high-throughput technology for alx1 cis*-regulatory analysis*

In order to accelerate the collection of *cis*-regulatory data, we employed a new high-throughput approach for rapid, parallel discovery and quantitative characterization of regulatory DNA sequence (Nam et al., 2010). This method permits the simultaneous introduction of multiple *cis*-regulatory expression constructs into the same batch of eggs. The activity of each individual reporter construct in the injected mixture is subsequently deconvolved by identification and quantification of its transcription product using sequence tags incorporated in the constructs, which act as unique "barcodes". The activity of each "barcoded" reporter is thus assayed independently in the nucleic acid extracted from the embryos by QPCR. This strategy enabled experimental measurement of the regulatory functions of multiple 0.5-2kb genomic DNA sequences, together with positive and negative controls, in each experiment. Control experiments in which all the construct tag vectors were driven by the same known active *cis*-regulatory module displayed subtle, i.e., less than 2-fold, variation in tag-specific expression (Supplementary Fig 2.1). While this amount of variation does not affect screening for weak or strong activator modules, it could interfere with quantitative detection of minor differences, for instance in assessing the effects of site mutations. To eliminate this source of variation, a normalization factor was obtained for each tag by averaging tag-specific activity over 5 repeated control experiments. These tag specific normalization factors were applied to all subsequent data obtained with the high-throughput tag system.

Scanning for functional non-coding sequence patches near the alx1 gene

To identify putative regulatory modules, we looked for sequence patches in the vicinity of the *alx1* gene that are conserved between *S. purpuratus* and *L. variegatus* genomes. Phylogenetic footprinting was carried out using Family Relations software (Brown et al., 2002) to compare *alx1* BAC sequences from the two species, using a 50 bp sliding window within which > 80% sequence identity was required. The region of overlap between the *Sp* and *Lv* BAC sequences was approximately 75 kb, and included all 5 *alx1* introns as well as 35 kb of upstream genomic DNA. This analysis identified 14 non-coding conserved sequence patches (labeled A-N) that lay within 25kb of exon1 (Figure 2.3a). The conserved patches ranged in size from 247 bp to 1735 bp and had an average size of 1 kb. These sequences were isolated by PCR and fused into a set of barcode tag vectors that contained the *alx1* basal promoter with the GFP coding sequence serving as the reporter (see Methods). Of all fragments tested (C through N), only fragment J generated levels of expression higher than background at 10 hpf (Figure 2.3b), which corresponds to the first peak of *alx1* expression. Fragments A and B, which are not included in Fig 2.3b, were independently tested and found to be inactive at both 10 and 24hpf.

To exclude the possibility of additional functional regulatory sequences that are non-conserved or only weakly conserved, an additional series of serially truncated BAC sequence fragments was tested. These were directly amplified from the *alx1* GFP BAC. These non-conserved sequences were labeled in lowercase (a-m) such that, for instance, element "a" lies between elements A and B. Figure 2.4a shows that reporter construct

I'→J, which includes nonconserved region i and conserved module J, was the shortest reporter

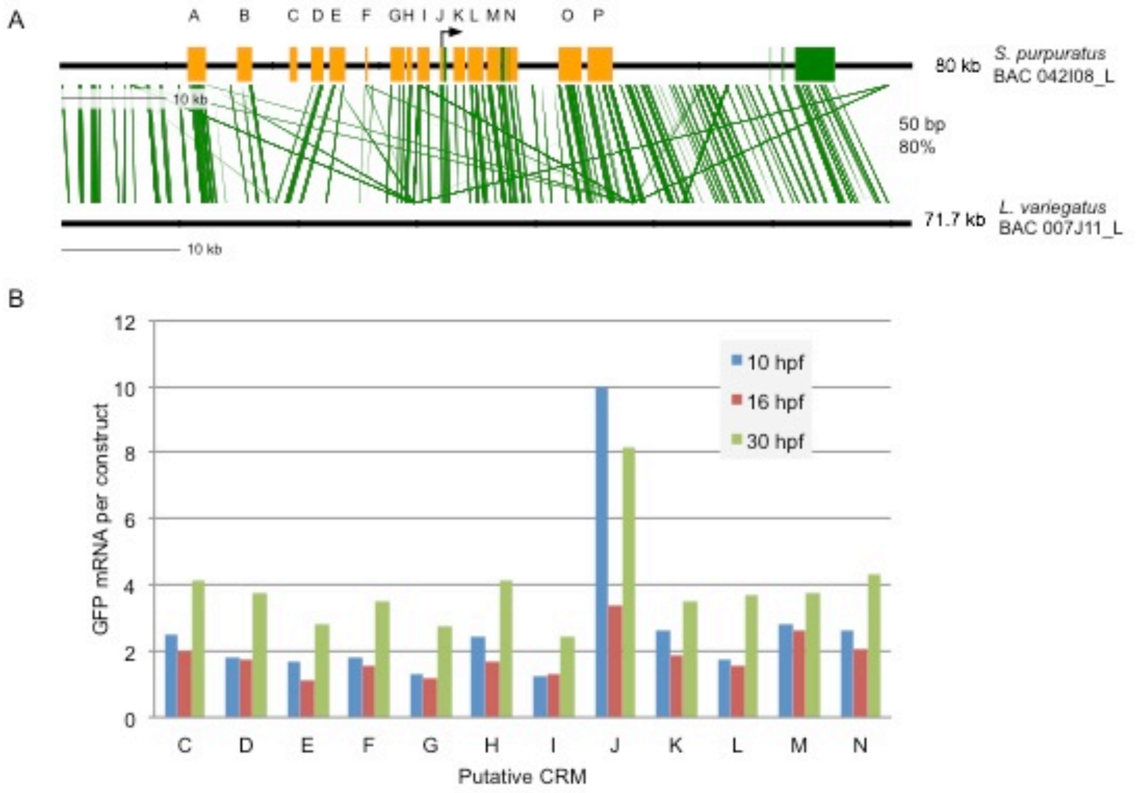


Fig 2.3. Phylogenetic footprinting and scanning of activity of cis-elements lying within 15kb of the *alx1* promoter. A) A 150 kb BAC (Lv_007J11_L) was isolated from an *L.v.* genomic library and sequenced. Phylogenetic footprinting was performed using seqComp (Brown et al., 2002) with a 50 bp/80% identity window. 14 conserved elements (labeled A-N) were found to lie between -22 kb and +8 kb relative to the first exon. B) 12 elements were examined for their ability to drive expression at 10, 16 and 30 hpf when fused to a GFP expression cassette containing the *alx1* basal promoter. Multiple constructs were measured independently using a tagged GFP technology and normalized for differences in the number of integrated genomic copies.

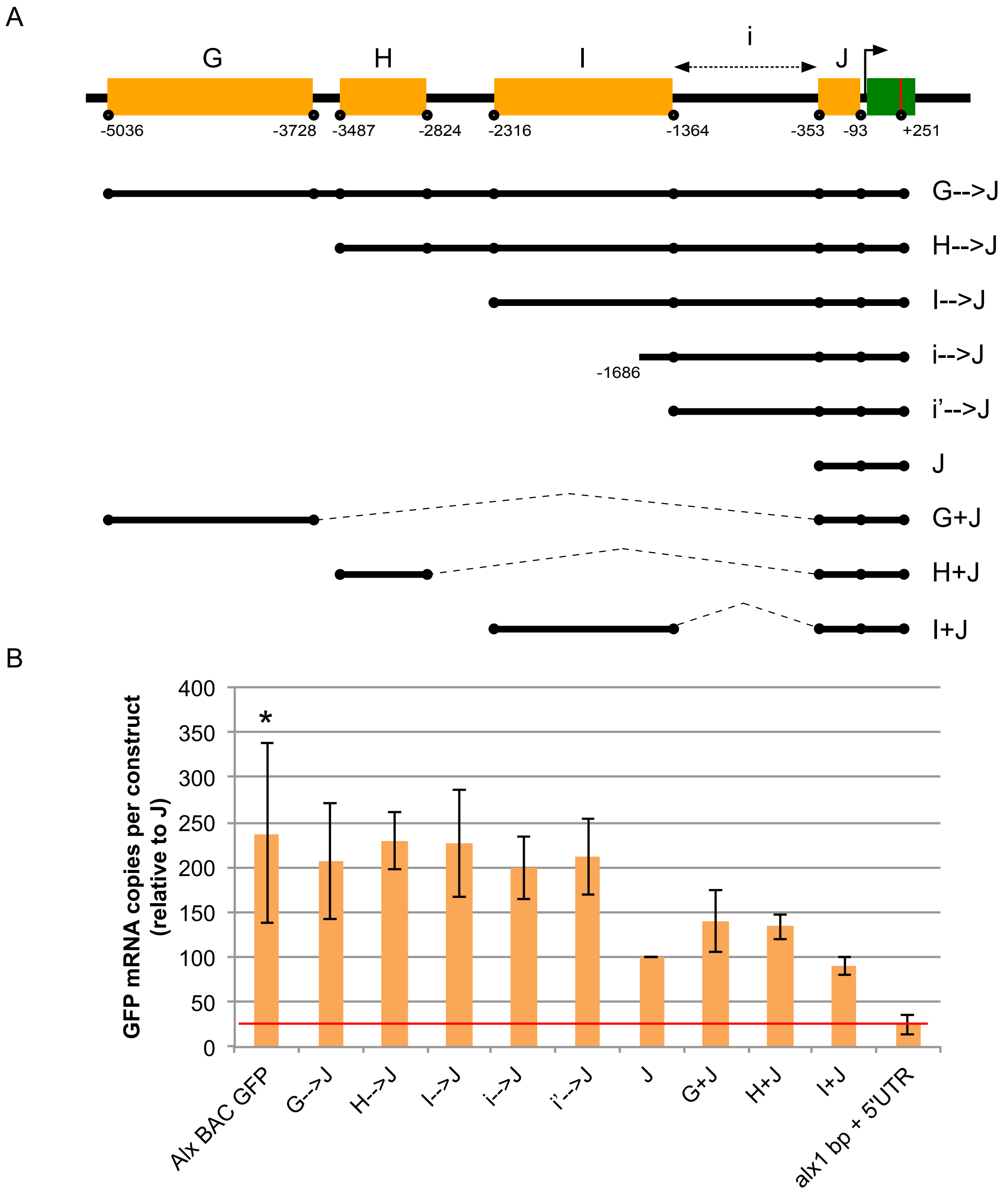


Figure 2.4. Serial Truncation of Alx1 GFP BAC reporter construct identifies a non-conserved, functional regulatory sequence. A) A series of tagged-GFP reporter constructs were generated to test the activity of regulatory sequences upstream of the promoter-proximal conserved module J B) qPCR data of injected reporter constructs was compared against activity of the full-length *alx1* GFP BAC at 11-12 hpf. Data is represented as normalized GFP expression relative to a reporter construct containing the J-module, *alx1* basal promoter, and *alx1* 5'UTR. Expression data was also normalized for tag-specific variation as described. Negative control expression level is marked by a red line.

capable of quantitatively matching *alx1* BAC GFP output levels per incorporated construct molecule (at 11.5 hpf). To be certain that region i contained no short conserved elements, a high-resolution Family Relations analysis with window size of 10 bp and 90% similarity was performed, but no regions of conservation were found (Supplemental Figure. 2.3). Comparison with the activity of J alone shows that inclusion of non-conserved region i materially boosts the output of J (though it is not capable of driving expression by itself), but no other additional upstream sequence further affected construct output (Figure. 2.4b).

Direct transcriptional repression of alx1 *by HesC*

As reviewed briefly above, expression of the initial tier of genes constituting the SM regulatory state, including *alx1*, is confined to the SM lineage by HesC repression everywhere else. Expression of these genes is permitted to occur in the micromere descendants because the initial gene of the double negative gate, *pmar1*, specifically prevents *hesC* transcription in these cells, where *pmar1* is activated soon after the lineage founder cells are born at 4th cleavage (Oliveri et a, 2002; 2003; Revilla et al, 2007). When *HesC* MASO is injected into fertilized eggs, *alx1* expression is up-regulated 4-7 fold (Revilla-i-Domingo et al., 2007). Taken together with the short time lag, 2.5-3 h, between *pmar1* and *alx1* activation, these results predict that HesC directly represses the *alx1* gene (cf. the kinetic study of gene cascades in this embryo; Bolouri and Davidson, 2003). Functional promoter-proximal Hesc binding sites of identical sequence have already been identified in two other genes that are activated coordinately with *alx1* and are also under control of the double negative gate, *tbrain* and *delta* (Smith and Davidson, 2008; Wahl et al., 2009). Putative HesC sites are also present near the promoter for the *ets1* gene, another double negative gate target gene (S. Damle, unpublished data). Thus we sought to determine if functional HesC binding sites exist in the accurately expressed J *alx1* minimal expression construct (Table 3.1). The *S. purpuratus hesC* gene is a member of the Hairy/E(spl) family of transcription factors (Howard-Ashby et al., 2006), which prefer class-B (CACGTG) or class-C (CACGCG) E-box sites (Fischer and Gessler, 2007). The repressive functions of Hairy/E(spl) members are mediated by their cofactor Groucho, an obligate transcriptional repressor, with which they specifically interact (Grbavec and Stifani, 1996; Paroush et al., 1994). *Hesc* is indeed expressed ubiquitously outside the SM lineage in the pre-hatching embryo, and its expression becomes more intense in the endoderm prior to hatching (Revilla-i-Domingo et al., 2007). A GFP-BAC reporter for the *hesC* gene corroborates this observation, showing stronger expression in presumptive endoderm at 24 hpf compared to ectoderm and NSM (Smith and Davidson, 2008).

Two class-C E-box sites in fact flank the TATA-box of the *alx1* J-construct. The proximal class-C site sits only 10bp downstream of the TATA box, while the distal site lies 47 bp upstream (Figure 2.5a). When both sites are mutated (Figure 2.5b), the J construct is upregulated 2-fold at 10-12 hpf and it expresses ectopically in ectoderm, veg1 and veg2 lineages at 18hpf (see “hesc-2xmut J” in Table1). This result is consistent with the network prediction that a ubiquitously expressed activator initiates *alx1* expression and that the interaction of Hesc at the proximal and distal binding sites is responsible for blocking *alx1* expression in all non-skeletogenic cells of the embryo. We complemented

Figure 5

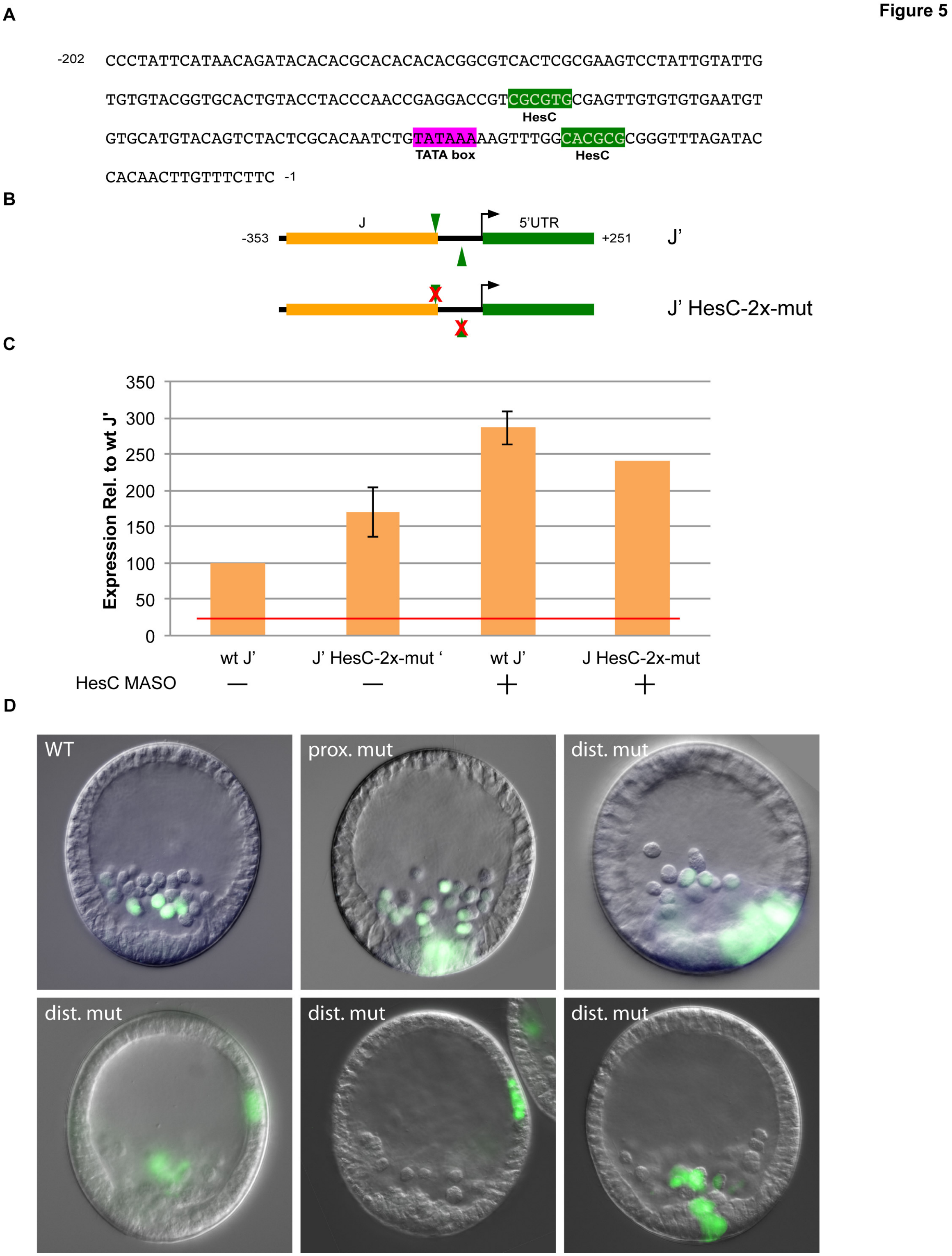


Fig 2.5. Effects of mutation of promoter-flanking Class-C bHLH sites on *alx1* expression. A) Sequence of the promoter-proximal region, showing two class-C bHLH binding sites (green highlight) flanking the TATAA box (purple highlight). B) Map of J' construct,

showing conserved J-module (orange), *alx1* 5'UTR (green) and class-C bHLH sites (green triangles) in antiparallel orientation (top) and map of mutant J' construct where both bHLH binding sites are mutated as described. C) QPCR assay of reporter GFP expression of HesC binding site mutant construct with and without *hesC* MASO. Expression is reported relative to wild-type J' construct shown in (B) and data is normalized for number of incorporated constructs and tag-specific variation (n=3) Negative control expression level is marked by a red line. D) Spatial expression pattern of BAC-GFP constructs whereby either the downstream (prox. mut) or upstream (dist. mut) HesC binding site is mutated. Top left, wild-type BAC; top middle, proximal-HesC site mutant BAC; top right and bottom row, distal-HesC site mutant BAC

this result with a *cis-trans* test, which shows that unlike the native J-construct, the double-mutant J construct is insensitive to co-injection with *hesc* MASO (Figure 2.5c).

Do these two HesC sites control spatial expression in the context of the whole gene as well as in the minimal construct? Perhaps the whole gene regulatory system might include some additional spatial control mechanism, so that although mutation of the HesC sites in J construct indeed produces ectopic expression, mutation of these two 6 bp sequences in the native gene might fail to derange normal spatial expression, which in context could be controlled by other interactions as well. Recalling that the genomic region 3' of the gene was not included in the conserved sequence scan (Figure 2.3a), such additional interactions could, for example, be mediated at an unexplored module located in this region. To examine such possibilities, we used BAC recombineering to remove each site from the complete *alx1* GFP BAC, and tested the spatial expression of the mutant BAC constructs. But the results were essentially the same as for the mutated J construct. In the complete context of the *alx1* GFP BAC, when either the proximal or distal HesC binding site is mutated, striking ectopic expression results, as illustrated in Figure 2.5d. Statistics collected for embryos injected with the mutant *alx1* BAC GFP reporters showed that, at 24 hpf for example, there was 4x more ectopic expression in NSM when the proximal HesC site alone is mutated; and when the distal HesC site is mutated there was 2x more expression in NSM, 5x more expression in the endoderm, and 2.5x more ectopic expression in the ectoderm, than in the parental BAC construct (Table 3.1). Taken together, these *cis*-regulatory results plus the *hesC* MASO data indicate that Hesc is in fact the direct input responsible for repressing *alx1* gene expression outside of the SM lineage up through mesenchyme blastula; that these HesC sites function to restrict expression to the SM lineage in the context of the whole gene; and that the genomic locus of the repressive input is specifically the two class-C E-boxes included in J construct that flank the promoter, both of which are necessary for complete repression. In their absence, expression is not properly restricted, so there is no effective additional spatial restriction system active in early development in this gene.

Cis-regulatory identification of the activator of early alx1 *gene expression*

Ets1 is thought to be to be a positive input to *alx1* (Ettensohn, 2003; Sharma and Ettensohn, 2010; Oliveri et al, 2008), and *ets1* mRNA of maternal origin is initially present at relatively high levels. The *ets1* gene is expressed zygotically before 12 hpf (Rizzo et al., 2006). Two putative Ets binding sites of the form ($^{C}/_{A}$GGAA) are present in the J sequence between -423 and -299, and three additional Ets binding sites in the I sequence between -1105 and -795 (Figure 2.6a, b). Mutation of the two J Ets sites within the context of the larger i→J construct decreased expression by over 50%, and deletion of these sites by about 3-fold (Figure 2.6c). In contrast, mutational analysis of the 3 sites within the more distal region i proved they are not strongly required. An additional more detailed series of deletions summarized in Supplemental Figure 2.5a displayed no additional cryptic sites in J.

An additional experiment demonstrates that Ets1 specifically interacts with *ets* binding sites identified in i→J. We measured expression of the construct of the 5-site ets mutant form of i→J in the presence of *ets1* MASO. If Ets1 is a direct activator of *alx1* then the activity of the 5-fold mutant should be insensitive to *ets1* MASO and should equal the level of expression of normal i→J in *ets1* MASO embryos, and this was the result obtained (Figure 2.6c).

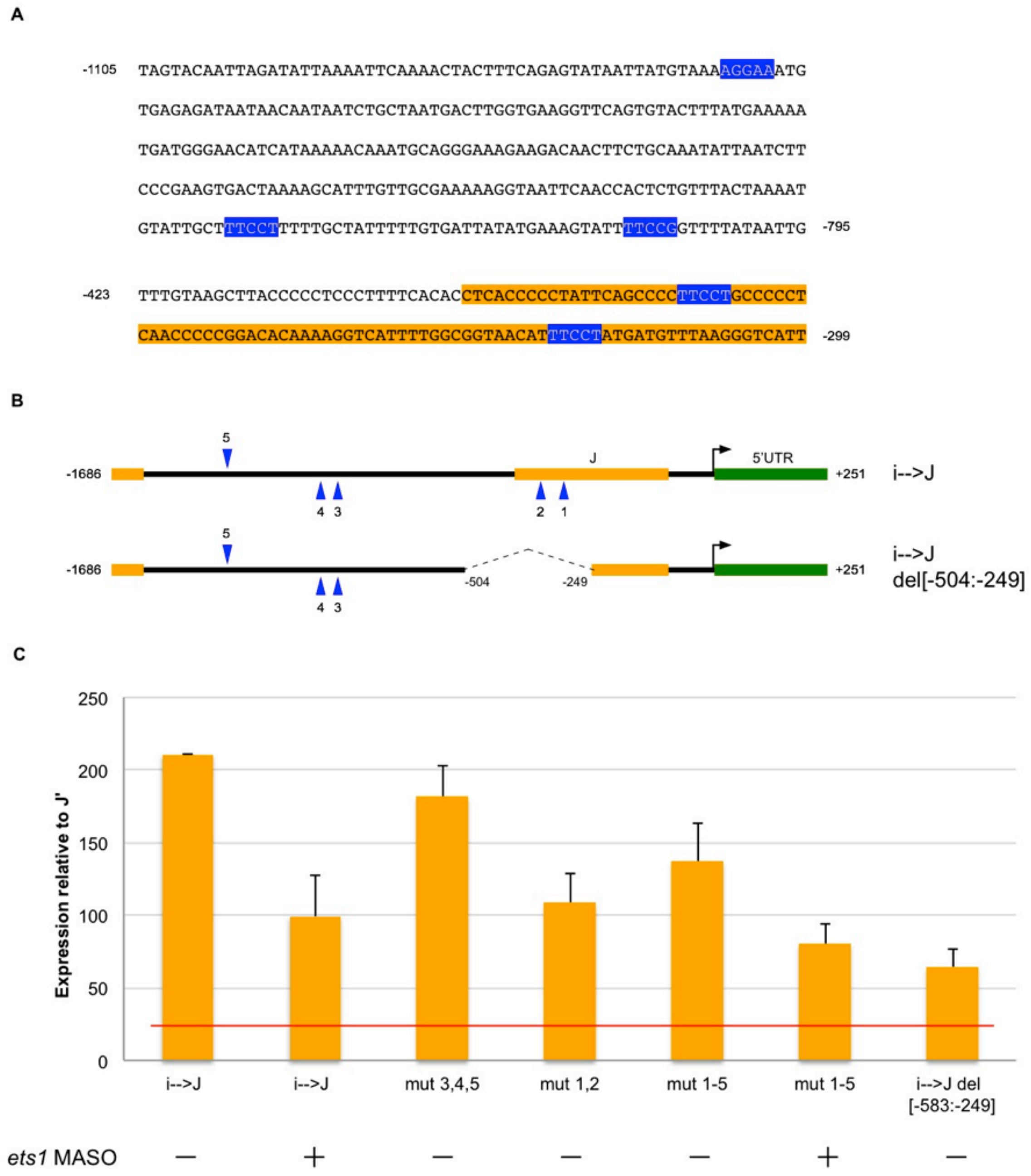


Fig 2.6. Ets1 sites in i→J are necessary for expression. A) Sequences within module i and J that contain Ets binding sites (blue highlight) of the form MGGAA. J-module sequence is highlighted in yellow. B) Map of i→J construct and deletion construct showing Ets1 binding sites (blue triangles). C) QPCR assay of reporter GFP activity of Ets1-binding site mutant and deletion constructs in the presence of *ets1* MASO at 11-12 hpf. Activity is reported relative to expression of wt construct i→J' shown in (B) and data is normalized for number of incorporated constructs and tag-specific variation (n=4). Negative control expression level is marked by a red line.

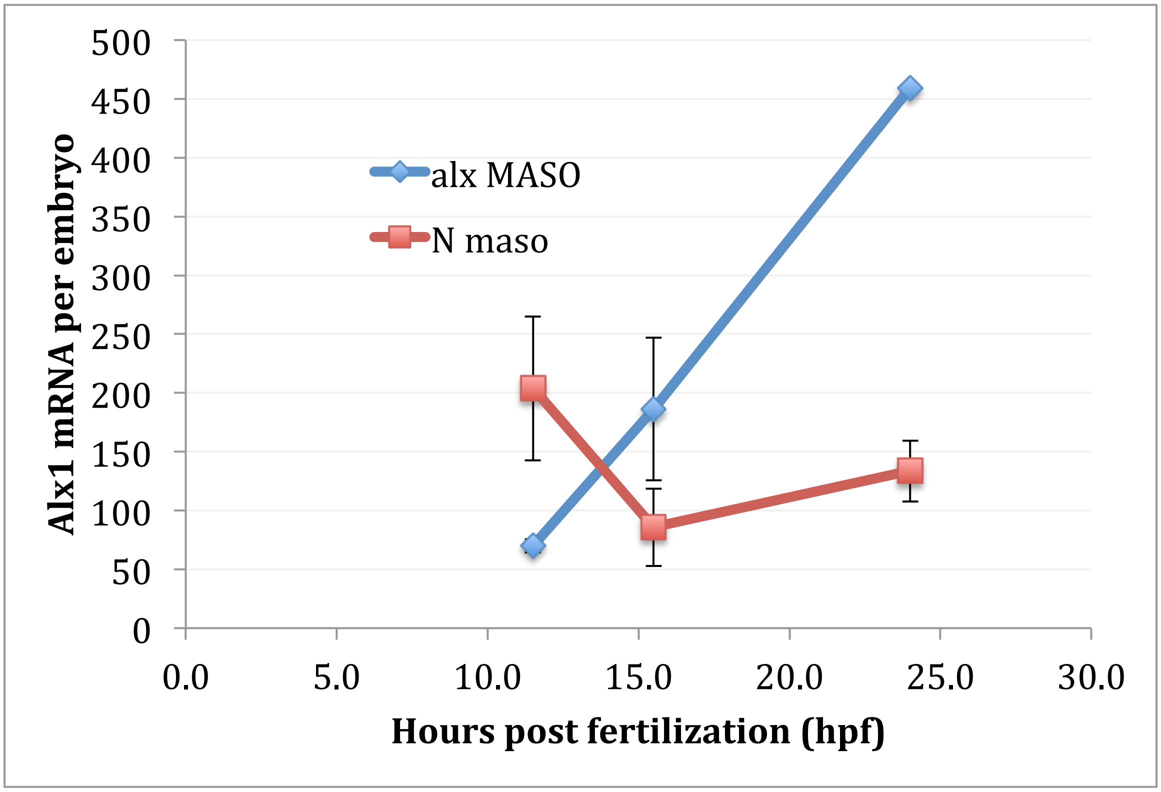
Alx1 modulates its own transcription

In earlier work Ettensohn et al. (2003) showed that alx1 is apparently repressed by its own gene product, since post-hatching embryos injected with *alx1* MASO displayed elevated levels of *alx1* transcript (Ettensohn et al., 2003). We carried out similar experiments but assayed the results by QPCR at the three key periods in the *alx1* expression time-course: the first expression peak at 11.5 hpf, the lowest point at 16 hpf and the late expression peak at 24 hpf (Figure 2.7a). MASO knockdown led to a 4-fold reduction in *alx1* at the first peak of expression, which was followed by an equal but opposite fold increase in expression at 16hpf and thereafter. These later results were consistent with the earlier post-hatching blastula measurements, but the early time point revealed that *alx1* also has a positive auto-regulatory input on its own expression during the dramatic early rise in transcript level. Because the i→J construct quantitatively recapitulates the early *alx1* expression profile (Figure 2.4b), it must contain the regulatory sequences responsible for both auto-activation and repression.

Homeodomain transcription factors bind to regulatory DNA via a helix-turn-helix motif that canonically recognizes AT-rich binding sites including the element TAAT. Members of the vertebrate class of Cart/Alx family contain a "Q50" Paired-type homeodomain that can dimerize cooperatively, and the dimer binds to a pair of palindromic half-sites that are separated by 3 bp, known as P3 sites (Wilson et al., 1993). Five putative monomeric Alx1 binding sites in module i→J were identified, but mutation of all of these sites collectively had no affect on construct expression levels (at 11.5 hpf; data not shown). We then looked for P3 sites of the form TAATNNNATTA. One such

P3 sequence exists within the 5'UTR and 3 others in region "i" (Figure 2.7c). Mutation of

all



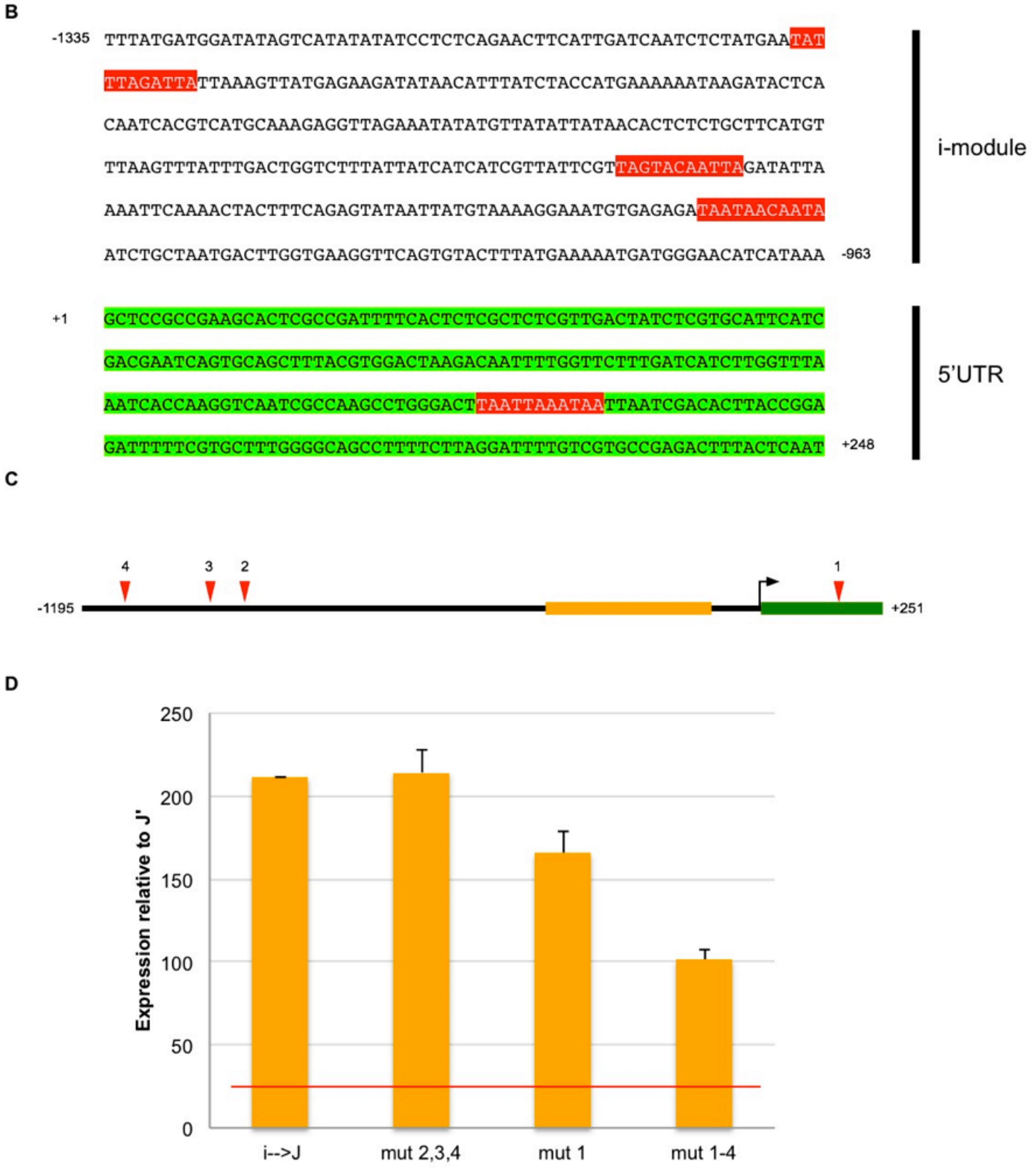


Figure 2.7. P3 sites in modules i and J' drive *alx1*. (A) qPCR timecourse of endogenous *alx1* mRNA in response to alx1 MASO (blue line) and control MASO (red line) injection. (B) Sequences within modules i and J that contain Alx1/Cart family binding sites of the form TAATNNNATTA. (C) Map of i→J construct showing Alx1 p3 sites (red triangles). (D) qPCR assay of reporter GFP activity of Alx1-binding site mutant constructs at 11-12hpf. Activity is reported relative to wild-type i→J expression and data is normalized for number of incorporated constructs and tag-specific variation (N=3). Negative control expression level is marked by a red line.

four of these sites indeed led to a > 50% drop in i→J activity at the 11.5 hpf expression peak (Figure 2.7d). Interestingly, the elimination of either the three distal sites in the 'i' region or mutation of the single site in the 5'UTR has only a small (or no) effect on reporter activity (Figure 2.7d). These results show that multiple *alx1* target sites are required for the normal level of early activation.

However, none of these four P3 sites are responsible for the apparent auto-repression at 16 hpf, since the 4-fold mutant shows the same expression profile as the wt construct, albeit with lowered levels of activity (Supplemental Fig 2.6a). It is this *alx1*-dependant repression that causes the precipitous decline in *alx1* transcript levels after 11.5 hpf. Thus when coinjected with *alx1* MASO, construct i→J is upregulated 2-3 fold at 16 hpf relative to controls, and endogenous *alx1* is up-regulated 3-4 fold (Ettensohn et al., 2003; Supplemental Figure 2.6B). We made a systematic attempt to scan for sequences responsible for *alx1*-dependant repression, testing a series of deletion constructs of the i→J reporter where consecutive 100-200bp sequences were removed. We calculated the ratio of expression of these reporters at 11.5 to 16 hpf to look for deletion constructs that exhibited derepression (i.e., which would display a ratio closer to 1). However, all constructs showed strong repression at 16hpf relative to 11 hpf (ratios > 2.5; Supplemental Figure 2.7). These results predict that while the auto-activation discovered here is direct, the apparent auto-repression is indirect, as we further consider below, in Discussion.

Alx1 *as both activator and repressor*

The evidence points clearly to the Janus-like behavior of the Alx1 transcription factor: it functions both as a repressor and an activator (on many down-stream genes as well as itself; Ettensohn et al, 2003; Oliveri et al, 2008). One possibility is that at lower levels of expression, Alx1 exists predominantly as a monomer, which acts as a transcriptional activator, whereas at high levels it dimerizes and becomes a repressor (or attracts a dimer-dependent co-factor which acts as a repressor). To test this, we synthetically generated an obligate dimer form of Alx1 by joining two Alx1 coding sequences with a linker encoding multiples of 4 glycines followed by a serine. Glycine-based linkers are commonly used for this purpose because they lack a β-carbon and therefore permit greater polypeptide flexibility. A serine interspersed between the glycine repeats acts to slow unfolding, thereby providing a useful amount of rigidity to the tether (Robinson and Sauer, 1998). The assumptions used to estimate the minimal length required to join two Alx1 monomers were: 1) that the monomers would be connected from N-terminus to C-terminus (see Figure 2.8a); 2) that Alx1 can be considered a spherical, globular protein with average density of 0.73cm^3/gm (Harpaz et al., 1994); 3) that the minimum tether distance should at least span the diameter of Alx1; and 4) that the N- and C- termini are not buried within the protein. Under these assumptions, the 440aa Alx1 protein would have a molecular weight of approximately 50kDa and a diameter of 48 Å. Given an average peptide unit length of 3.8 Å (Iwakura and Nakamura, 1998), we chose a linker sequence repeated 3 times (G4Sx3) to obtain a tether with total length of 57 Å. This tether length should be sufficiently long to permit dimerization on antiparallel strands under most protein configurations, except for cases where one

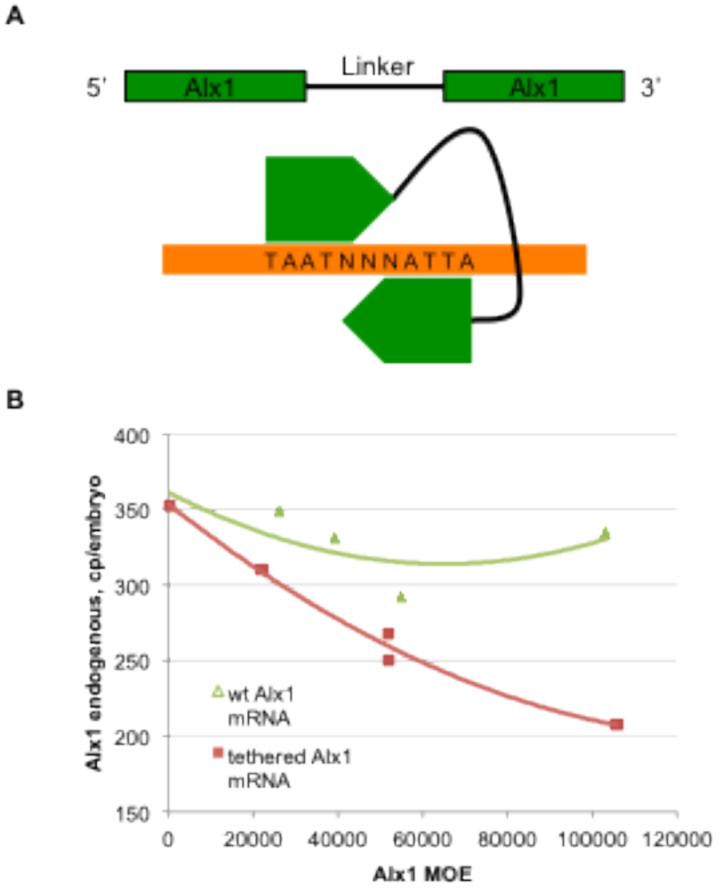


Fig 2.8. Alx1 MASO has dual effects on *alx1* transcription. (A) (top) An obligate dimer form of Alx1 was generated by inserting sequence encoding a G4Sx3 (glycinex4 + serine) tether in between two *alx1* coding sequences. This construct was fused to endogenous 5'UTR and 3'UTR sequence and in-vitro transcribed and injected into fertilized embryos. (bottom) A model showing Alx1 homodimer binding in an antiparallel orientation to target P3 sites. Linker length is estimated at 57 Angstroms and diameter of a single Alx1 monomer is estimated to be roughly 48 Angstroms. (B) Effect on endogenous *alx1* mRNA levels in response to a titration of *alx1* mRNA or *tethered alx1* mRNA injection. Abscissa is injected mRNA levels measured at 11-12hpf and ordinate reflects levels of endogenous *alx1* mRNA.

monomer’s N-terminus is very far from the other monomer’s C-terminus. We then analyzed the N- and C-terminal regions of Alx1 and found they contain several hydrophilic residues, suggesting that the ends of Alx1 are indeed exposed, and therefore should be able to be tethered without significantly impairing tertiary structure. The tethered fusion protein was named Alx1-G4Sx3-Alx1. mRNA encoding this protein was injected in a titration experiment into fertilized eggs, and endogenous *alx1* transcript was measured at 11.5 hpf by QPCR. The results (Figure 2.8b) showed unequivocally that the obligate dimer represses *alx1* by over 2-fold compared to the monomer. Interestingly, in addition to acting as a repressor of *alx1* expression, the Alx1-G4Sx3-Alx1 dimer is a more potent activator of the differentiation *alx1* target gene *msp130L* (Supplemental Figure 2.8). Taken together, these results show that Alx1 dimerizes to perform auto-regulatory functions of both polarities, but while the P3 palindromic double half-site promotes dimerization on the DNA, higher concentration, for which the obligate tether provides a surrogate, may result in spontaneous formation of dimers that have repressive activity on some genes.

DISCUSSION

Using recombinant BAC reporter knock-ins, and short regulatory expression constructs derived from the BAC, we have solved the *cis*-regulatory system responsible for all aspects of pre-gastrular *alx1* expression, spatial, quantitative, and temporal. Target binding sites and their inputs have been identified. This work has addressed several issues that until now were outstanding: first, what are the specific genomic regulatory features that link the *alx1* gene into the SM double-negative gate control system; second, what are the activators and genomic regulatory features responsible for driving *alx1* expression in the large micromeres; third, what regulatory controls explain the "peak and valley" temporal kinetics of *alx1* message. An outcome of this study is a further elaboration of the GRN subcircuit wiring surrounding the SM double negative gate, as summarized in Figures 9a-d.

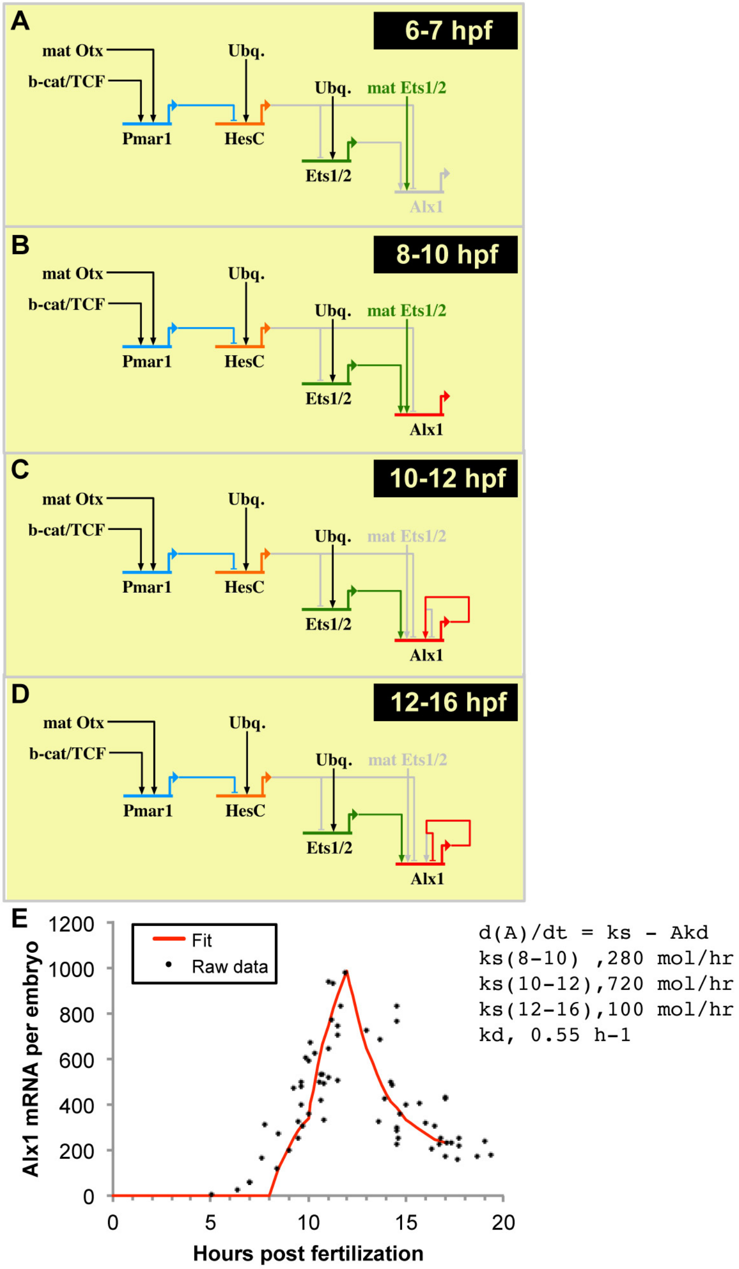


Fig 2.9. Network architecture controlling *alx1* expression in large micromeres and resultant kinetics. A-D) Inputs are labeled in order of the relative order of their effects on alx1 transcription. A) at 6-7 hpf, *pmar1* clears *hesc* from the large micromeres, B) at 8-10 hpf *ets1/2* initiates *alx1* transcription, C) from 10-12 hpf Alx1 protein autoactivates *alx1,* D) at 12 hpf Alx1 begins to function as an autorepressor E) The kinetics of *alx1* transcription from Figure 1 were fit to the differential equation d(A)/dt = k_s – Ak_d (Davidson, 1986; Ben-Tabou de-Leon and Davidson, 2009) using the following parameter values: k_d is 0.55 h^{-1}, and k_s is the synthesis rate which varies as follows: from 8-10 hpf, k_s = 280 molecules per hour (mol/hr); from 10-12 hpf, k_s =720 mol/hr; and from

12-17 hp, k_s=100 mol/hr. These transcriptional rates are well within the limits of sea urchin transcriptional machinery at 15°C. Considering the most rapid rate of *alx1* transcription, 720 molecules/hr, and given there are 8 large micromeres containing 16 copies of the *alx1* gene, the initiation rate per gene copy is 45 transcripts/hr or 1 transcript every 80 seconds. This value is roughly 8 times slower than the computed maximal initiation rate at 15C of 1 transcript every 9 seconds (Davidson, 1986; pp.144-145)

Kinetics of alx1 *expression*

As shown in Figure 2.9e, the temporal *alx1* expression data can be fit very well on the basis of the conclusions drawn in this work. The assumptions that were used to generate the kinetics shown in Figure 2.9E are as follows: (1) There is an initial rate of gene expression which obtains from the activation of the gene at about 7.5 hpf until enough time has elapsed for the *alx1* mRNA to accumulate and be translated to effective levels (expected to be 2-3 h; Bolouri and Davidson, 2003), here taken as until 10 hpf. (2) Thereupon a sharp increase in the synthesis rate occurs, due to auto-activation, which from the time-course accumulation data of Figure 2.1, and the mutation data of Figure 2.7e, is about twice the initial rate, and this enhanced rate obtains until ~ 11 hpf. (3) The peak of expression is due to the transformation of the Alx1 gene product into a repressor. Since none of the target sites for Alx1 dimers in constructs displaying this repression mediate repression, the repressive effect is indirect, and a reasonable but untested hypothesis is that the repression is actually directed at *ets*, the obligate driver of *alx1*. (4) After the accumulation peak the effect of the repression is to decrease the rate of synthesis, which falls to a constant rate obtaining to the end of the period considered, here 17hpf. (5) The turnover rate is intrinsic to the mRNA and is constant throughout. It can be estimated from the declining phase of the expression time course, and considering the trough at ~ 16 hpf as a steady state, the synthesis rate after the peak can then be calculated. The parameters used are shown in the inset in Figure 2.9E; for mathematical approach see legend. The point to be made here is that the processes defined experimentally in this work, particularly (1-3) in the foregoing, suffice to explain the time course of *alx1* expression. Their significance is straightforward. The auto-activation mechanism serves to drive up the transcript concentration much more rapidly than would otherwise be possible (compare for example the relatively leisurely accumulation of *tbrain* mRNA as shown by Wahl et al; *tbr* is also a target of the double negative gate, but lacks the auto-activation device). But every positive feedback needs to be damped sooner or later or the product accumulates exponentially. The conversion of Alx1 to a repressor, which as suggested by the synthetic experiment of Figure 2.8 is probably due to a concentration-dependent dimerization mechanism, self-limits the auto-activation. This results in a decline in *alx1* transcript and eventually a new, lower, steady state obtains pending the late phase of increased transcription (Figure 2.1), which we did not analyze here.

Control of alx1 *expression by the double negative gate*

Functional *cis*-regulatory evidence now directly substantiates the double negative gate architecture proposed earlier for three target genes, *tbr*, *delta*, and now *alx1*, and likely *ets1* operates by means of the same types of HesC target sites as have been demonstrated to control spatial expression of these genes. Of the regulators constituting the definitive SM regulatory state, this leaves only *tel* yet to be validated by *cis*-

regulatory evidence as a direct double negative gate target gene. The evidence is of the same form for all three genes: the predicted HesC target sites are found to be present, and when mutated in expression constructs are demonstrated to be absolutely required to confine expression to the SM domain, just as predicted from extensive earlier data (Oliveri et al., 2002; 2003; Revilla et al, 2007; Oliveri et al, 2008; Smith et al, 2008). The evidence is perhaps strongest for *alx1* and *tbr*, as in both cases the complete genomic landscape in which the gene is embedded, carried in a large BAC recombinant, was subjected to functional *cis*-regulatory analysis, so there is little possibility that a missing regulatory module might have escaped attention. The dramatic effect of mutating only the two 6 bp HesC target sites of the *alx1* gene in the 129,000 bp BAC precludes the possibility suggested by Sharma and Ettensohn (2010), that there is another, different spatial control system confining *alx1* expression to the skeletogenic lineages (in *Lytechinus variegatus*). A species difference could account for their results (this would not be the first such discovered), but more likely, it is just not sufficient to draw conclusions on *cis*-regulatory apparatus from immunocytology and other indirect observations when only specific *cis*-regulatory measurements can provide definitive evidence. The kinetics of *alx1* expression add to the picture when taken together with the *cis*-regulatory evidence that the driver of *alx1* gene expression is Ets1, which conforms perfectly to the conclusions also drawn from ets1 MASO studies (Ettensohn et al, 2003; Oliveri et al., 2008; Sharma and Ettensohn, 2010). At least in *S. purpuratus* the *alx1* gene does not become active immediately after the micromeres are born, when the first cohort of SM-specific genes are activated, i.e., *pmar1* and *blimp1* (Oliveri et al, 2002; Revilla et al, 2007; Smith and Davidson, 2008), but only about 2h later, at 7.5hpf (Revilla et al, 2007; this work); the useable Ets1 in the micromeres is evidently the zygotic product of the newly activated *ets1* gene (see below), and both *ets1* and *alx1* require *pmar1* to have been expressed in order for them to be transcribed normally. We examined the possibility that the localized maternal regulator Otx, which is a required driver of *pmar1*, could also provide an input into *alx1* even though this would not be consistent with the timing of *alx1* activation, but found that mutation of all the possible Otx target sites in our expression constructs has no effect whatsoever on expression levels (data not shown). This is in contrast with results of just such mutation experiments on the *pmar1* and *blimp1* genes, which are indeed controlled by Otx drivers (Smith and Davidson, 2008; Smith et al, 2007). In summary, *cis*-regulatory evidence surrounding the SM double negative gate now extends from the *pmar1* and *hesC* genes (Smith and Davidson, 2008) to the *tbr* (Wahl et al, 2009), *delta* (Revilla et al, 2007; Smith et al, 2008), and *alx1* genes immediately downstream, and as a result of this study both the spatial and temporal particularities of *alx1* expression have now been incorporated in a consistent explanatory framework based ultimately in genomic regulatory sequence design.

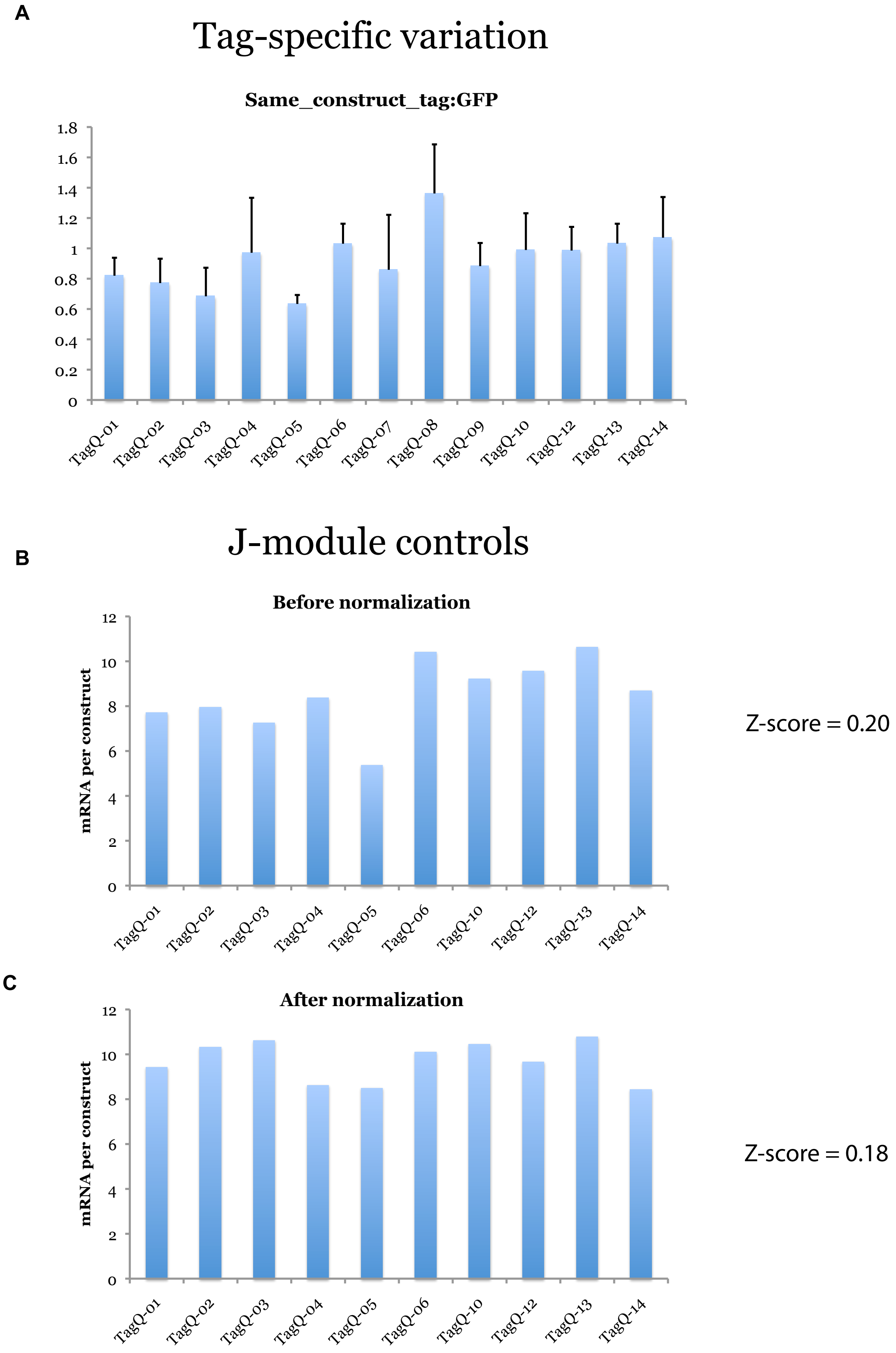
The elegance of the SM double negative gate architecture

In early development spatial expression is controlled at least as much by spatially confined repressors, coupled with wide spread activators, as by locally confined activators (for reviews, Davidson, 2001; 2006). The SM double negative gate is an especially elegant device for initiating spatially confined embryonic gene expression. Its parts list includes the primordially localized micromere inputs Tcf/βcatenin and Otx (Oliveri et al, 2008); one or two zygotically expressed ubiquitous transcriptional activators (unknown); the first zygotically activated spatially confined positive regulator in the system, *ets1*; a zygotically activated, transcriptionally confined repressor, *pmar1*; a zygotically activated but ubiquitously expressed second repressor, *hesC;* and a set of downstream SM regulatory state genes of which we are here concerned mainly with *alx1*, *tbr*, and *delta*. As discussed by Peter and Davidson (2009), the double negative gate operates globally in a Boolean fashion, in that it not only causes expression of its downstream targets in the confined SM domain, but accomplishes active repression of the same genes everywhere else in the embryo. The discovery that Ets1 is the direct positive driver of *alx1*, just as it is of *delta* and *tbr*, while the *ets1* gene itself is also subject to HesC repression and a target of the double negative gate, adds a beautiful wiring feature to the regulatory architecture. This is captured in Figs. 2.9a-d. Here we see, in "View from the Nucleus" BioTapestry models, the sequence of events. First *pmar1* is specifically activated in response to the primordially localized inputs. Then 1.5-2 h later, an unknown activator that is evidently ubiquitous appears in the embryo, and turns on the *hesC* gene, everywhere except in the SM lineage, where the *hesC* gene is dominantly repressed by the *pmar1* gene product (for high-resolution relative expression kinetics, see Revilla et al, 2007; Materna et al, 2010). Driven by perhaps the same ubiquitous activator, the *ets1* gene is also turned on, but, in the exact Boolean opposite of *hesC*, only in SM cells, not elsewhere because the HesC repressor is now elsewhere. The SM-specific, zygotically expressed Ets1 now serves as the driver of the other target genes, *alx1*, *tbr*, and *delta*. This timing, and indeed the fact that the *hesC* double negative gate is what determines SM-specific expression of all these target genes probably means that (the globally distributed) maternally encoded *ets1* mRNA is not their driver, or else these genes could be activated all over the embryo, before the *hesC* gene is activated. When the spatial performance of the double negative gate is destroyed, by *hesC* MASO, or by global expression of *pmar1*, global expression of the other target genes results (Oliveri et al., 2002;2003; Revilla et al., 2007): the reason is that now expression of their *ets1* driver becomes global. In view of the foregoing we feel the ectopic expression of the Hesc binding site mutant BAC is due not to Ets1/2 in the remainder of the embryo but either to another ets factor that binds the ets target sites (like *ets4*) or an ubiquitous activator that gives a small amount of activity beyond the inputs described in this work. Finally, the relative timing of the successive states of the double-negative gate is a crucial aspect of the spatial control mechanism mediated by the hard-wired genomic regulatory circuitry shown in Figures 2.9a-d.

Supplementary Table 2.1 Binding site mutations of alx1 cis-regulatory architecture

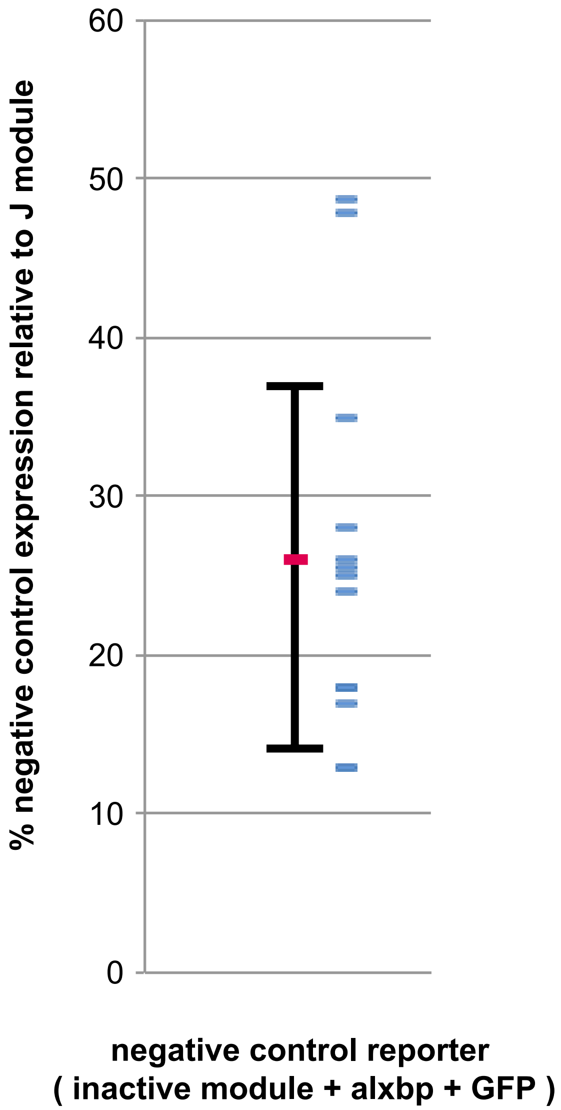
Binding Site Name	WT sequence	Mutated/Deleted sequence	Postion*	Orientation
Alx site 1	ACTTAATTAAATAATTAAT	ACTTAgcTAgATAcagAAT	153	
Alx site 2	AGATAATAACAATAATC	AGATgcTAACAcgAATC	-1036	
Alx site 3	CGTTAGTACAATTAGAT	CGTctGTACAAcgAGAT	-1105	
Alx site 4	GAATATTTAGATTATTA	GAATgcTTAGAgcATTA	-1276	
Ets site 1	CATAGGAAATG	CATtGaAtATG	-323	-
Ets site 2	GGCAGGAAGGG	GGCAtagcGGG	-373	-
Ets site 3	AACCGGAAAAT	AACCttcAAAT	-812	-
Ets site 4	AAAAGGAAAGC	AAAAttcAAGC	-849	-
Ets site 5	AAAAGGAAATG	AAAAttcAATG	-1051	+
Proximal Hesc site	TGGCACGCGCGG	TGGacatatCGG	-37	+
Distal Hesc site	TCGCACGCGACG	TCGacatatACG	-102	-

* Position relative to start site of transcription

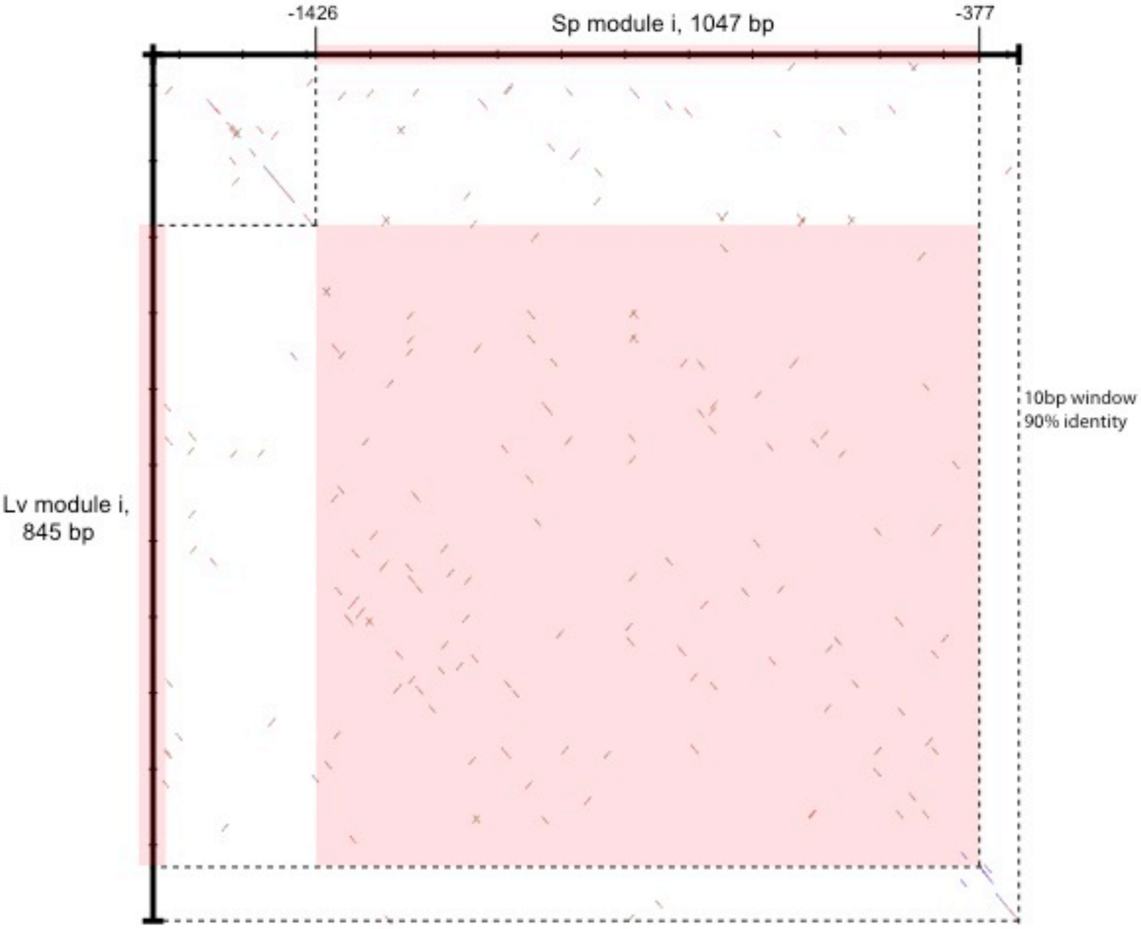


Supplementary Figure 2.1. Tag-specific variation of tagged GFP reporter constructs. A)

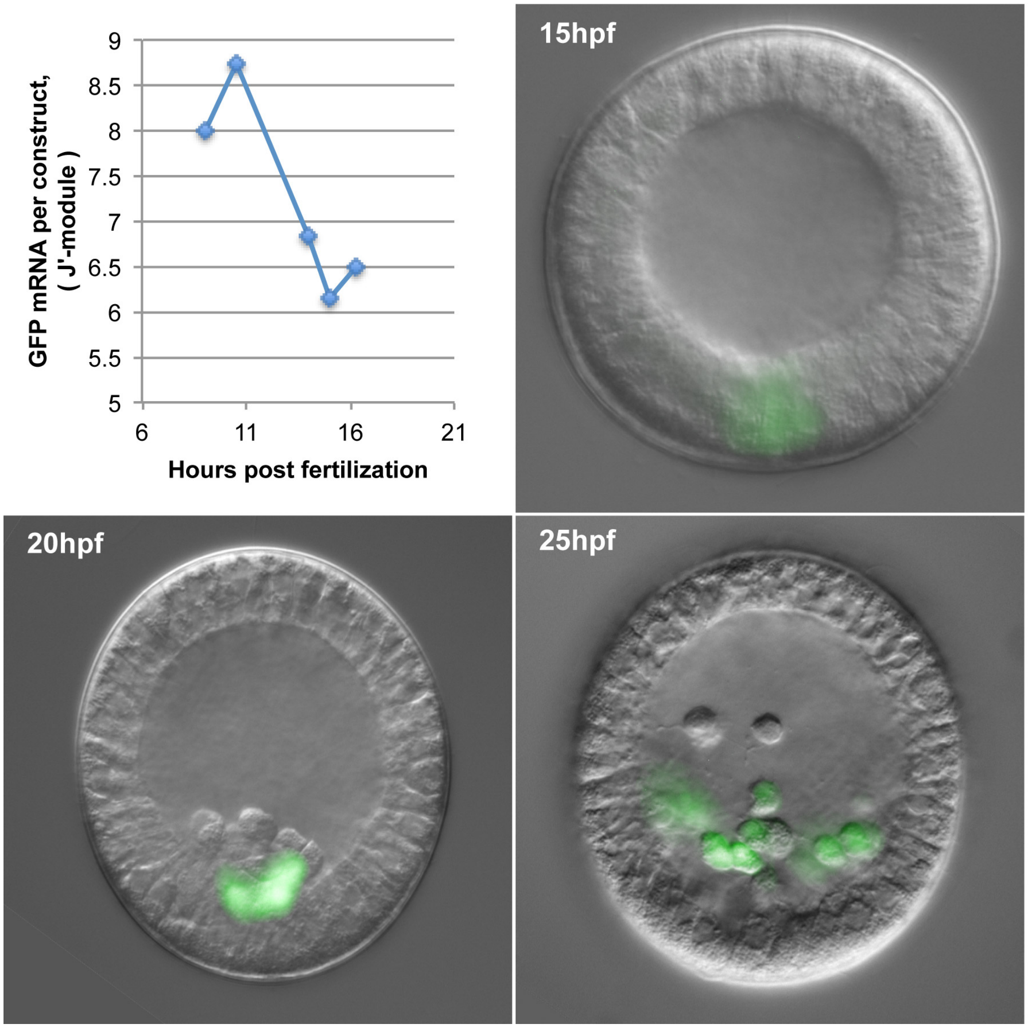
Tag-specific expression of reporter constructs, all of which contain identical J' modules. N=5. B) Similar J'-module control experiment before normalization. C) J'-module control experiment after normalizing for tag-specific variation. Construct values for each tag in (B) were divided by tag-specific variation in (A).



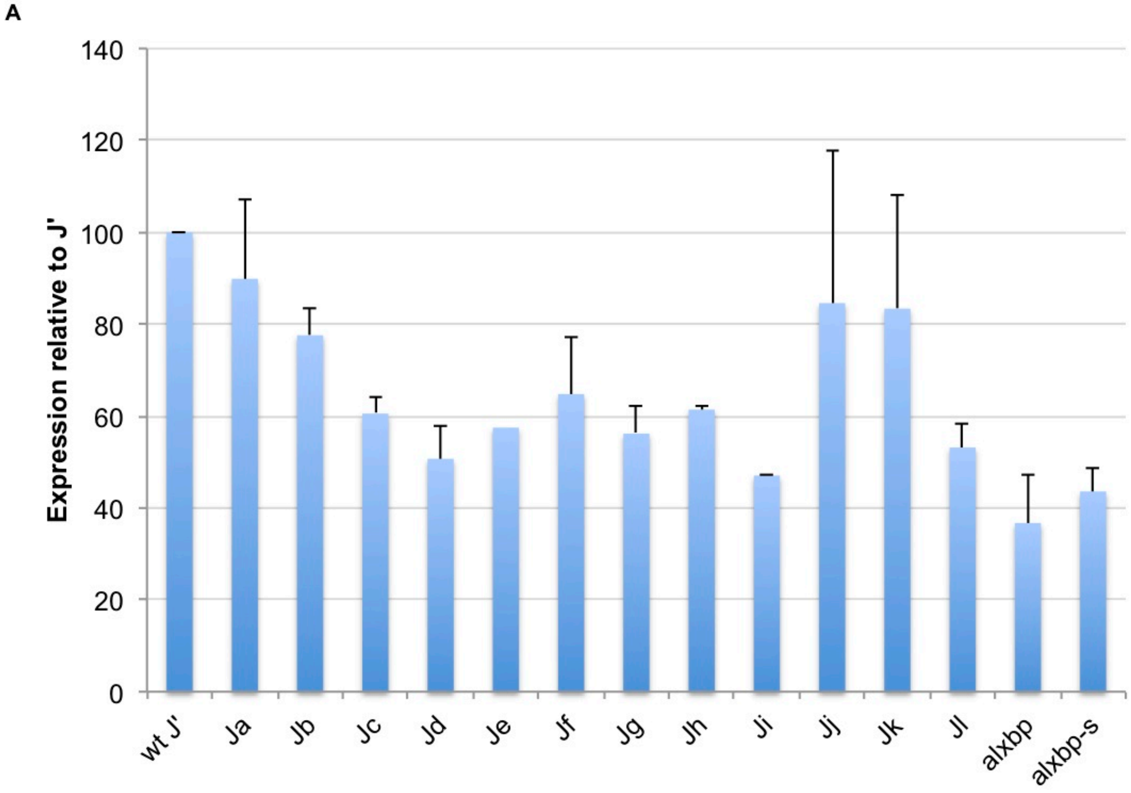
Supplementary Figure 2.2. Negative control levels relative to J' expression. Negative controls were generated by fusing alx basal promoter and 5'UTR to non-functional cis-elements derived from Alx1 BAC sequence. (N=14 negative controls)



Supplementary Figure 2.3. Dotplot of non-conserved ‘i’ module. Module ‘i’ was defined as the region between conserved modules I and J in *S.p.* and *L.v. alx1* BAC sequences. A dot-plot analysis was performed comparing 10bp windows with 90% sequence identity between *S.p.* and *L.v.* sequence. Pink box denotes non-conserved ‘i’ region. Matching sequences are colored at the nucleotide level according to the following key: G, purple; C, blue; A, green; T, red.



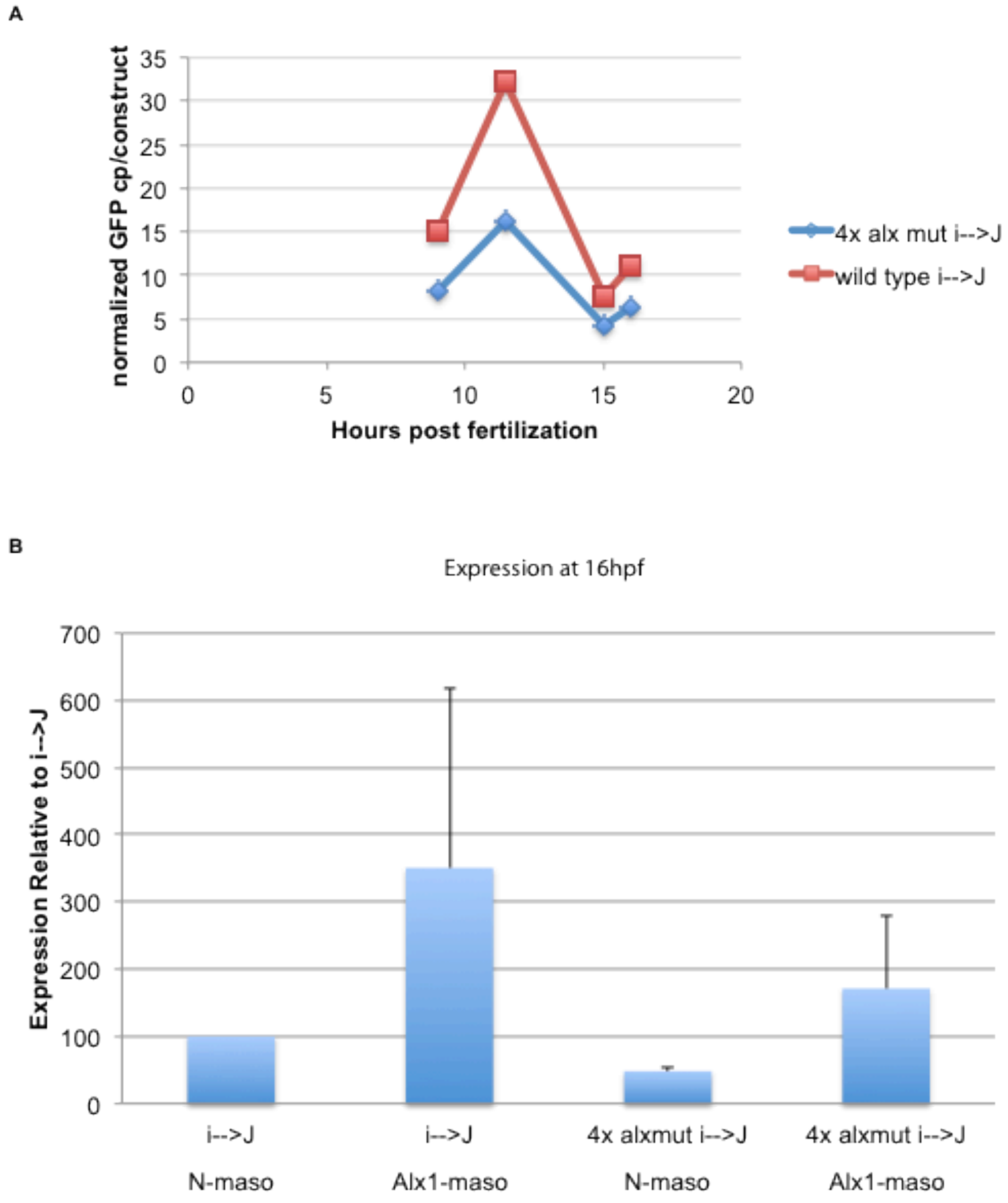
Supplementary Figure 2.4. Expression of J' construct. Embryos injected with GFP reporter driven by J' were imaged for fluorescence at the indicated timepoints.



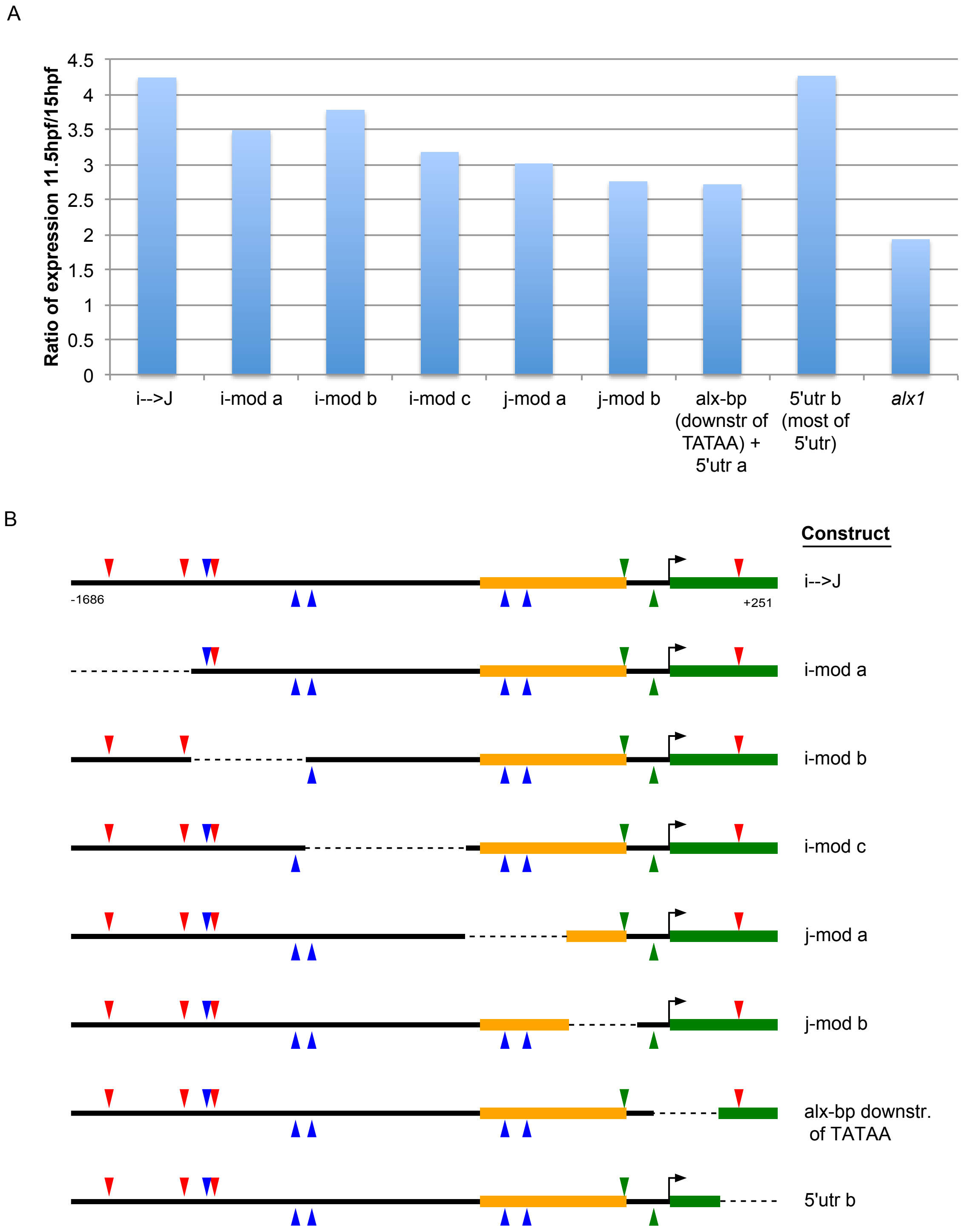
B

```
> J sequence

  wtJ-->         Ja-->                         Jb--> Jc -->
CTTACCCCCTCCCTTTTCACACCTCACCCCCTATTCAGCCCCTTCCTGCCCCCTCAACCCCC
           Jd-->                     Je-->              Jf-->
GGACACAAAAGGTCATTTTGGCGGTAACATTTCCTATGATGTTTAAGGGTCATTTCTTTTGA
            Jg-->     Jh-->                       Ji-->
GAAACGCGAAAGCACCTGATTTGCACTCTTGACCAATGACCGTGCCCGAAGCCCAGCGGTGT
                     Jj-->                Jk-->
ATAATAGCACAAAAAGGATGCCGCTCGCCCTATTCATAACAGATACACACGCACACACACGG
                                 Jl-->
CGTCACTCGCGAAGTCCTATTGTATTGTGTGTACGGTGCACTGTACCTACCCAACCGAGGAC
              alxbp-->               alxbp-s-->
CGTCGCGTGCGAGTTGTGTGTGAATGTGTGCATGTACAGTCTACTCGCACAATCTGTATAAA

AAGTTTGGCACGCGCGGGTTTAGATACCACAACTTGTTTCTTCGCTCCGCCGAAGCACTCGC

CGATTTTCACTCTCGCTCTCGTTGACTATCTCGTGCATTCATCGACGAATCAGTGCAGCTTT

ACGTGGACTAAGACAATTTTGGTTCTTTGATCATCTTGGTTTAAATCACCAAGGTCAATCGC

CAAGCCTGGGACTTAATTAAATAATTAATCGACACTTACCGGAGATTTTTCGTGCTTTGGGG

CAGCCTTTTCTTAGGATTTTGTCGTGCCGAGACTTTACTCAATATTG
```

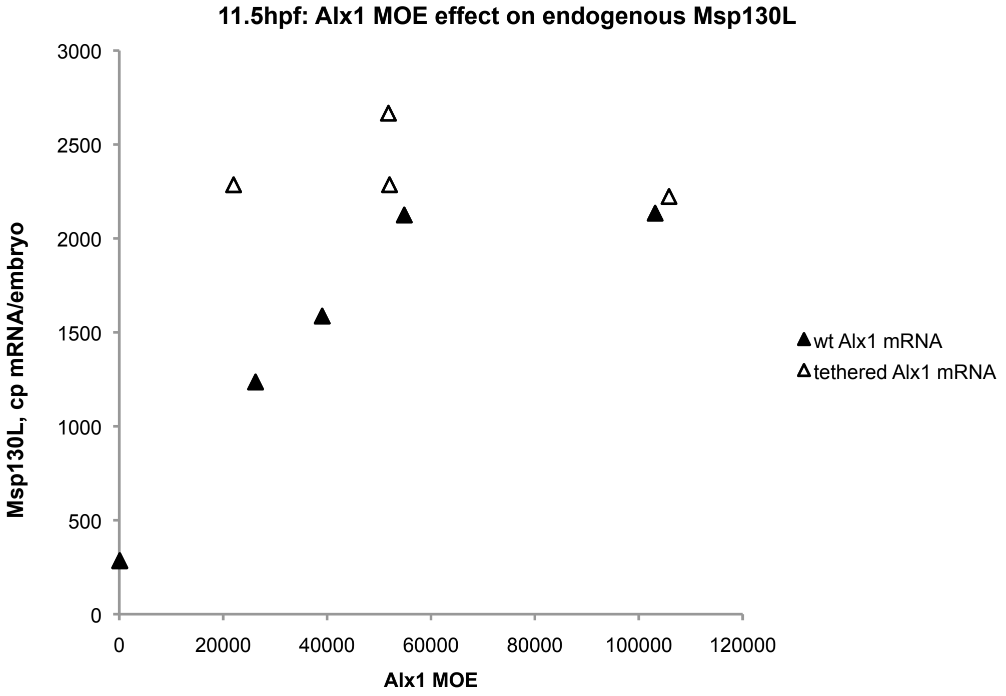
Supplementary Figure 2.5. Serial Truncation of J' construct identifies enhancers in upstream portion of the J-module. A) A series of J' truncation constructs were injected into embryos and assayed for tagged-GFP expression at 11.5hpf B) Map of 5'-most position of each serial truncation. Ets1 sites are labeled in blue highlight. TATAA box is labeled in yellow highlight. Start site of transcription is underlined.



Supplementary Figure 2.6. Effect of Alx1 MASO on i’→J’ construct and on Alx1-binding site mutant i’→J’. A) Expression timecourse comparing wt i→J reporter and 4x-alx1-site mutant. B) Embryos injected with i→J and with a mutated version of i→J containing 4 alx1-binding site mutations were assayed for expression in the presence of Alx1 MASO or N-MASO at 16hpf. (n=2)



Supplementary Figure 2.7. Systematic deletion of i'→J' construct does reveals no Alx1-mediated autorepression domains. A) A series of deletion constructs (deletions roughly 200-300bp) were injected into embryos and assayed for tagged-GFP expression at 11.5hpf. B) Map of deletion constructs showing deleted regions as dashed line. Alx1 binding site, red triangle; Ets1 binding site, blue triangle; HesC binding site, green triangle.



Supplementary Figure 2.8. Alx1 homodimer is a potent activator of skeletogenic differentiation battery gene, Msp130L. The transcriptional effect on endogenous *msp130L* by injection of *alx1-G4Sx3-alx1* mRNA at 4 different concentrations was compared to wild-type *alx1* mRNA at the first peak of *alx1* expression (11.5 hpf).

REFERENCES

Ben-Tabou de-Leon, S., Davidson, E.H., 2009. Modeling the dynamics of transcriptional gene regulatory networks for animal development. Dev Biol 325, 317-328.

Beverdam, A., Meijlink, F., 2001. Expression patterns of group-I aristaless-related genes during craniofacial and limb development. Mech Dev 107, 163-167.

Bolouri, H., Davidson, E.H., 2003. Transcriptional regulatory cascades in development: initial rates, not steady state, determine network kinetics. Proc Natl Acad Sci U S A 100, 9371-9376.

Brown, C.T., Rust, A.G., Clarke, P.J.C., Pan, Z., Schilstra, M.J., De Buysscher, T., Griffin, G., Wold, B.J., Cameron, R.A., Davidson, E.H., Bolouri, H., 2002. New computational approaches for analysis of cis-regulatory networks. Developmental Biology 246, 86-102.

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

Court, D.L., Sawitzke, J.A., Thomason, L.C., 2002. Genetic engineering using homologous recombination. Annu. Rev. Genet. 36, 361-388.

Davidson, E.H., 1986. Gene Activity in Early Development. Academic press, Orlando.

Davidson, E.H., Rast, J.P., Oliveri, P., Ransick, A., Calestani, C., Yuh, C.-H., Minokawa, T., Amore, G., Hinman, V., Arenas-Mena, C., Otim, O., Brown, C.T., Livi, C.B., Lee, P.Y., Revilla, R., Schilstra, M.J., Clarke, P.J.C., Rust, A.G., Pan, Z., Arnone, M.I., Rowen, L., Cameron, R.A., McClay, D.R., Hood, L., Bolouri, H., 2002. A provisional regulatory gene network for specification of endomesoderm in the sea urchin embryo. Developmental Biology 246, 162-190.

Ettensohn, C.A., Illies, M.R., Oliveri, P., 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.

Fischer, A., Gessler, M., 2007. Delta-Notch--and then? Protein interactions and proposed modes of repression by Hes and Hey bHLH factors. Nucleic Acids Res 35, 4583-4596.

Gao, F., Davidson, E.H., 2008. Transfer of a large gene regulatory apparatus to a new developmental address in echinoid evolution. Proc Natl Acad Sci USA 105, 6091-6096.

Grbavec, D., Stifani, S., 1996. Molecular interaction between TLE1 and the carboxyl-terminal domain of HES-1 containing the WRPW motif. Biochem Biophys Res Commun 223, 701-705.

Harpaz, Y., Gerstein, M., Chothia, C., 1994. Volume changes on protein folding. Structure 2, 641-649.

Hobert, O., 2002. PCR fusion-based approach to create reporter gene constructs for expression analysis in transgenic C. elegans. BioTechniques 32, 728-730.

Howard-Ashby, M., Materna, S.C., Brown, C.T., Chen, L., Cameron, R.A., Davidson, E.H., 2006. Gene families encoding transcription factors expressed in early development of Strongylocentrotus purpuratus. Dev Biol 300, 90-107.

Iwakura, M., Nakamura, T., 1998. Effects of the length of a glycine linker connecting the N-and C-termini of a circularly permuted dihydrofolate reductase. Protein Eng 11, 707-713.

Lee, P.Y., Nam, J., Davidson, E.H., 2007. Exclusive developmental functions of gatae cis-regulatory modules in the Strongylocentrorus purpuratus embryo. Developmental Biology 307, 434-445.

Nam, J., Dong, P., Tarpine, R., Istrail, S., Davidson, E.H., 2010. Functional cis-regulatory genomics for systems biology. Proc Natl Acad Sci USA 107, 3930-3935.

Oliveri, P., Carrick, D.M., Davidson, E.H., 2002. A regulatory gene network that directs micromere specification in the sea urchin embryo. Developmental Biology 246, 209-228.

Oliveri, P., Davidson, E.H., McClay, D.R., 2003. Activation of pmar1 controls specification of micromeres in the sea urchin embryo. Developmental Biology 258, 32-43.

Oliveri, P., Tu, Q., Davidson, E.H., 2008. Global regulatory logic for specification of an embryonic cell lineage. Proc Natl Acad Sci USA 105, 5955-5962.

Paroush, Z., Finley, R.L., Jr., Kidd, T., Wainwright, S.M., Ingham, P.W., Brent, R., Ish-Horowicz, D., 1994. Groucho is required for Drosophila neurogenesis, segmentation, and sex determination and interacts directly with hairy-related bHLH proteins. Cell 79, 805-815.

Qu, S., Li, L., Wisdom, R., 1997. Alx-4: cDNA cloning and characterization of a novel paired-type homeodomain protein. Gene 203, 217-223.

Revilla-i-Domingo, R., Minokawa, T., Davidson, E.H., 2004. R11: a cis-regulatory node of the sea urchin embryo gene network that controls early expression of SpDelta in micromeres. Developmental Biology 274, 438-451.

Revilla-i-Domingo, R., Oliveri, P., 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 USA 104, 12383-12388.

Rizzo, F., Fernandez-Serra, M., Squarzoni, P., Archimandritis, A., Arnone, M.I., 2006. Identification and developmental expression of the ets gene family in the sea urchin (Strongylocentrotus purpuratus). Dev Biol 300, 35-48.

Robinson, C.R., Sauer, R.T., 1998. Optimizing the stability of single-chain proteins by linker length and composition mutagenesis. Proc Natl Acad Sci U S A 95, 5929-5934.

Sharma, T., Ettensohn, C.A., 2010. Activation of the skeletogenic gene regulatory network in the early sea urchin embryo. Development 137, 1149-1157.

Smith, J., Davidson, E.H., 2008. Gene regulatory network subcircuit controlling a dynamic spatial pattern of signaling in the sea urchin embryo. Proc Natl Acad Sci USA 105, 20089-20094.

Wahl, M., Hahn, J., Gora, K., Davidson, E., 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. Developmental Biology. 2009 Nov 15;335(2):428-41

Warming, S., Costantino, N., Court, D.L., Jenkins, N.A., Copeland, N.G., 2005. Simple and highly efficient BAC recombineering using galK selection. Nucleic Acids Res 33, e36.

Wilson, D., Sheng, G., Lecuit, T., Dostatni, N., Desplan, C., 1993. Cooperative dimerization of paired class homeo domains on DNA. Genes & Development 7, 2120-2134.

CHAPTER 3

Confocal Quantification of Cis-Regulatory Reporter Gene Expression in Live Sea Urchin Embryos

Sagar Damle*, Bridget Hanser*[†], Eric H. Davidson*, and Scott E. Fraser*[†‡]

*Division of Biology, [†]Beckman Institute, [‡]Division of Engineering and Applied Science,
California Institute of Technology, Pasadena, CA 91125

(Published in *Developmental Biology* 299 (2006) 543–550.)

ABSTRACT

Quantification of GFP reporter gene expression at the single-cell level in living sea urchin embryos can now be accomplished by a new method of confocal laser scanning microscopy (CLSM). Eggs injected with a tissue-specific GFP reporter DNA construct were grown to gastrula stage and their fluorescence recorded as a series of contiguous Z-section slices that spanned the entire embryo. To measure the depth-dependent signal decay seen in the successive slices of an image stack, the eggs were coinjected with a freely diffusible internal fluorescent standard, rhodamine dextran. The measured rhodamine fluorescence was used to generate a computational correction for the depth-dependent loss of GFP fluorescence per slice. The intensity of GFP fluorescence was converted to the number of GFP molecules using a conversion constant derived from CLSM imaging of eggs injected with a measured quantity of GFP protein. The outcome is a validated method for accurately counting GFP molecules in given cells in reporter gene transfer experiments, as we demonstrate by use of an expression construct expressed exclusively in skeletogenic cells.

Keywords: confocal laser scanning microscopy, GFP, sea urchin *cis*-regulation

INTRODUCTION

Developmental *cis*-regulatory analysis is carried out by injecting into eggs constructs in which the *cis*-regulatory DNA is associated with a reporter gene, whereupon reporter gene expression can be determined at given embryonic stages (Revilla-i-Domingo et al., 2004). There are a variety of choices of reporter available for sea urchin embryos, depending on the desired measurement. The expression of such *cis*-regulatory constructs can be estimated quantitatively by using a reporter gene encoding an enzyme the activity of which is determined in homogenates of transgenic embryos; or the quantity of reporter mRNA can be determined directly by QPCR (Yuh et al., 1998; Arnone et al., 2004; Revilla-i-Domingo et al., 2004). On the other hand, to assess spatial reporter gene expression requires the use of fluorescent reporters such as GFP (Chalfie et al., 1994), or whole mount in situ hybridization (WMISH) of the reporter mRNA can be carried out, irrespective of what the reporter encodes (Yuh et al., 2004). But these are either/or propositions: a major limitation, in every experimental embryonic system, is that either quantitative or spatial information regarding reporter gene expression can be derived from a given embryo, but not both. In situ hybridization, for example, retains spatial information but sacrifices quantitative information. The amount of message available for in situ hybridization differs for each target sequence, and the interactions between probe and target are not one-to-one, since a single message molecule may bind one to several labeled probes in an unpredictable manner. Alternatively, QPCR can accurately detect only a few transcripts per cell, (e.g., Lee and Davidson, 2004; Yuh et al., 2005), but at cost of destruction of the sample and the loss of any spatial information. Current experimental methods for *cis*-regulatory analysis thus require reporter expression to be measured spatially and quantitatively on separate samples in any given experiment.

Here we report a solution to this general problem in which the activity of GFP reporters is measured quantitatively in any desired cell(s) using confocal laser scanning microscopy (CLSM). Standard CLSM offers the possibility of quantitative imaging of thin optical sections (on the order of microns) in living specimens, but quantitation is confounded in thick samples, such as embryos, due to optical distortions and internal light scatter. While striking images can be collected with CLSM throughout an intact embryo, there is a significant loss of signal at depths exceeding 10-15 microns. A simple depth correction is not adequate, as different tissue types have different optical properties, varying the severity of signal loss. Recently we described in these pages a method by which absolute numbers of GFP molecules can be deduced from CLSM image stacks in whole sea urchin embryos (Dmochowski et al., 2002). To obtain an accurate measure of depth-dependent loss of signal, a freely-diffusible synthetic dye, Texas Red (TR)-labeled dextran, was injected into the eggs. Recovery of red fluorescence was then used to normalize total fluorescence in deep optical sections to that of their shallower counterparts. This approach laid the foundation for quantitative photomicroscopy, but it had yet to be extended to the measurement of reporter gene activity in individual cells.

The expression system used in the present work was a sea urchin *tbrain* (*tbr*) BAC GFP knockin; *tbr* is a regulatory gene expressed exclusively in skeletogenic mesenchyme, which functions in the GRN in the process of setting up the skeletogenic regulatory state upstream of expression of biomineralization genes (Oliveri and Davidson, 2004). We demonstrate the reproducible measurement of the number of molecules of

GFP protein produced by this construct in individual skeletogenic cells of living transgenic embryos.

MATERIALS AND METHODS

Sea urchin embryos and microinjections

Strongylocentrotus purpuratus gametes were prepared for microinjection as described (McMahon et al., 1985). De-jellied eggs were rowed onto injection dishes coated with a 1% protamine sulfate solution and fertilized with diluted sperm. Injected embryos were grown at 14°C and imaged 0.5, 24 or 52 h postfertilization (hpf). Because *S. purpuratus* embryos hatch from their fertilization membranes and begin to swim at 18 hpf, for imaging, those grown to 24 or 52 hpf were immobilized using a solution of .0025% poly-L-lysine in filtered seawater.

The injection solutions contained 12.5% glycerol, 120 mM KCl, 0.5 μg/μl rhodamine dextran (MW = 3 kD), a freely-diffusing small molecule used as an internal fluorescent standard. Embryos were coinjected with either a Tbrain-GFP BAC DNA construct or purified epGFP protein (BD Biosciences, molecular weight 30 kD), and depth-dependent decay of GFP fluorescence was corrected using the corresponding rhodamine-dextran intensity. Tbrain-GFP-BAC reporter was constructed using the BAC-recombination method (Yu et al., 2000). In this construct, the first exon of the coding sequence for the *tbrain* gene is replaced by that of a GFP-mutant, S65T. This mutant also enhances the translation efficiency of the GFP mRNA. Prior to injection, the BAC construct was linearized with AscI and purified on a CL-4B column.

Injection solutions had a total volume of 10 μl. Tbrain BAC constructs were 140 kb in length and injected at a concentration of 25 ng/μl (~160 molecules/pl). Purified rEGFP protein (MW 30,000 Daltons, BD Biosciences/Clonetech) was injected at a concentration of 500 ng/µl (10^7 molecules/pl). No carrier DNA was used in any microinjection solution.

Imaging with CLSM.

To maintain *S. purpuratus* embryos at an optimum culture temperature, an aluminum stage fitted with a Peltier cooling device was attached to the universal stage adapter of an Axiovert 100 M inverted microscope configured for CLSM (LSM 5 PASCAL, Zeiss).

The sea urchin egg is roughly spherical and has a diameter of 80 µm, whereas the post-gastrula embryo reaches a diameter of nearly 100 microns along its shortest axis, the animal/vegetal axis. Spherical aberrations are caused primarily by the difference between the refractive index of the immersion medium and the mounting medium. A lens displaying spherical aberration will not be able to focus all of the rays from the source of light into the detector. This can lead to a substantial loss of signal, which becomes a severe problem when scanning deep sections within a sample. To minimize this and other sources of signal decay, a C-Apochromat 40X water-lens (working distance of 290 µm) with large numerical aperture (1.2 NA) was used. Dmochowski *et al*. (Dmochowski et al., 2002) noted mild edge effects in the form of fluorescence peaks at both shallow and deep slices that are attributed to out of plane fluorescence and laser scatter. That our data show no such effects may be partly attributed to our choice of optics. The C-Apochromat 40X lens used for scanning uses an immersion medium, water, whose index of refraction closely matches that of the sample medium, seawater. Imaging under this setup, we were able to nearly eliminate spherical aberration, significantly increasing detection sensitivity in deep slices. Test experiments showed that a coverglass-correction-collar setting of 0.16 is least sensitive to spherical aberration when imaging the embryo mounted in sea water.

For all imaging experiments, we employed the same settings to simplify the comparison of results. The pinhole was set to 120 µm (1.5 Airy unit), and detector gains were set at 800 V for both red and green channels. The same laser power (argon ion, 488 nm 11% power, and helium-neon, 543 nm 80% power,) was used for all measurements. Emitted fluorescence from GFP and rhodamine was detected through a LP505 and LP560 filter, respectively. Embryos were imaged at scan speed 8 (1.76 µs per voxel, 1 sec per section) in 1.4 µm-thick z-sections for a total of 70-90 slices, and CLSM image data were stored as two separate stacks of images (one for each channel). The section dimensions were 230 microns in length and width and 1.2 microns thick. Fluorescence was recorded as a square 16-bit image with edge length of 512 pixels. As such, the dimensions of each pixel in an image represent a cellular volume 0.45 microns by 0.45 microns and 1.2 microns deep, or 0.24 cubic microns (0.24 femtoliters), the minimal unit of volume measured.

QPCR of GFP mRNA in injected embryos.

One hundred injected embryos were lysed at the 40 h (mid-gastrula) stage and total RNA was collected using a Qiagen RNeasy Micro kit. RNA was reverse-transcribed into cDNA with random hexamer primers and the Taqman RT kit (Applied Biosystems). QPCR was performed using primers for GFP and for a ubiquitously expressed reference gene, *SpZ12* (Wang et al., 1995).

Data Processing.

The confocal images for each embryo were recorded as 16-bit tif files and processed using the free software package ImageJ (Abramoff et al., 2004) and a custom-made ImageJ plugin, LSM Intensicor (written by Dr. Bridget Hanser). Each image set was comprised of two image stacks, one measuring GFP fluorescence and the other measuring rhodamine fluorescence. Each image in a stack was square, containing 512 pixels spanning 230 µm along each edge.

An image set was processed as follows: (*i*) Both GFP (488nm excitation) and rhodamine (543 nm excitation) stack images were despeckled using a 3x3 square median filter. (*ii*) A Gaussian filter was applied to each image within a stack. Filters with the following kernels: [1] (no Gaussian filter), [0.84, 1, 0.84] (filter = 1 pixel radius) and [0.249, 0.707, 1.000, 0.707, 0.249] (filter = 2 pixel radii) were used during the analysis. A threshold was set equal to the mean background fluorescence of each despeckled slice (140, 16-bit pixel units) and the image average intensity above background was then calculated. An array of voxels with red intensity exceeding the threshold was created contiguous with eight surrounding voxels (forming a 3x3 square) within the same horizontal plane, as described (Dmochowski et al., 2002). The thresholded average intensity value for each slice in a stack was normalized to the mean thresholded fluorescence of the shallowest slice. The coefficients of normalization constitute the depth-profile for the stack. (*iii*) A Gaussian filter was then applied to the green stack and the depth-profile was used to compensate for depth-dependent loss of GFP fluorescence. (*iv*) Background fluorescence, calculated from the dark space surrounding each embryo was subtracted. (*v*) Background embryo fluorescence (calculated from fluorescence of uninjected embryos) was subtracted from each measured embryo or cell.

RESULTS

The method and its validation.

In previous work we had developed a method for correction of depth-dependent loss of signal intensity in blastula stage sea urchin embryos (Dmochowski et al., 2002), and our present approach (Fig. 3.1) builds upon this. As before, the experimental embryos are prepared by injection of eggs with both a diffusible red fluorescent dye and the *cis*-regulatory expression construct which is to be the object of the study. The eggs are cultured to the desired stage, and an image stack obtained by confocal microscopy. In capturing the total fluorescence of an embryo, confocal microscopy produces a series of images, each image representing an individual horizontal slice, or optical section, through the sample. The ensuing data processing steps are as follows (see Materials and methods for details).

The images within the stack were first passed through a rank order ("despeckling" or median) filter that removes spurious pixels, that is, single pixels in which thermal noise creates bright puncta, or in which drop-out creates dark puncta. In a median filter the intensity of each pixel is replaced by the median of the intensities of the pixel and its eight neighbors. This significantly reduces the contribution of noise to the final analysis; typically, the overall fluorescence intensity is reduced by 20 units per square micron (Fig. 3.1*I*, *J*). A threshold equal to the mean background intensity is then set, and a mean intensity of all pixels above threshold is calculated for each image in the red channel. This constitutes the information on which the depth profile is based. After despeckling,

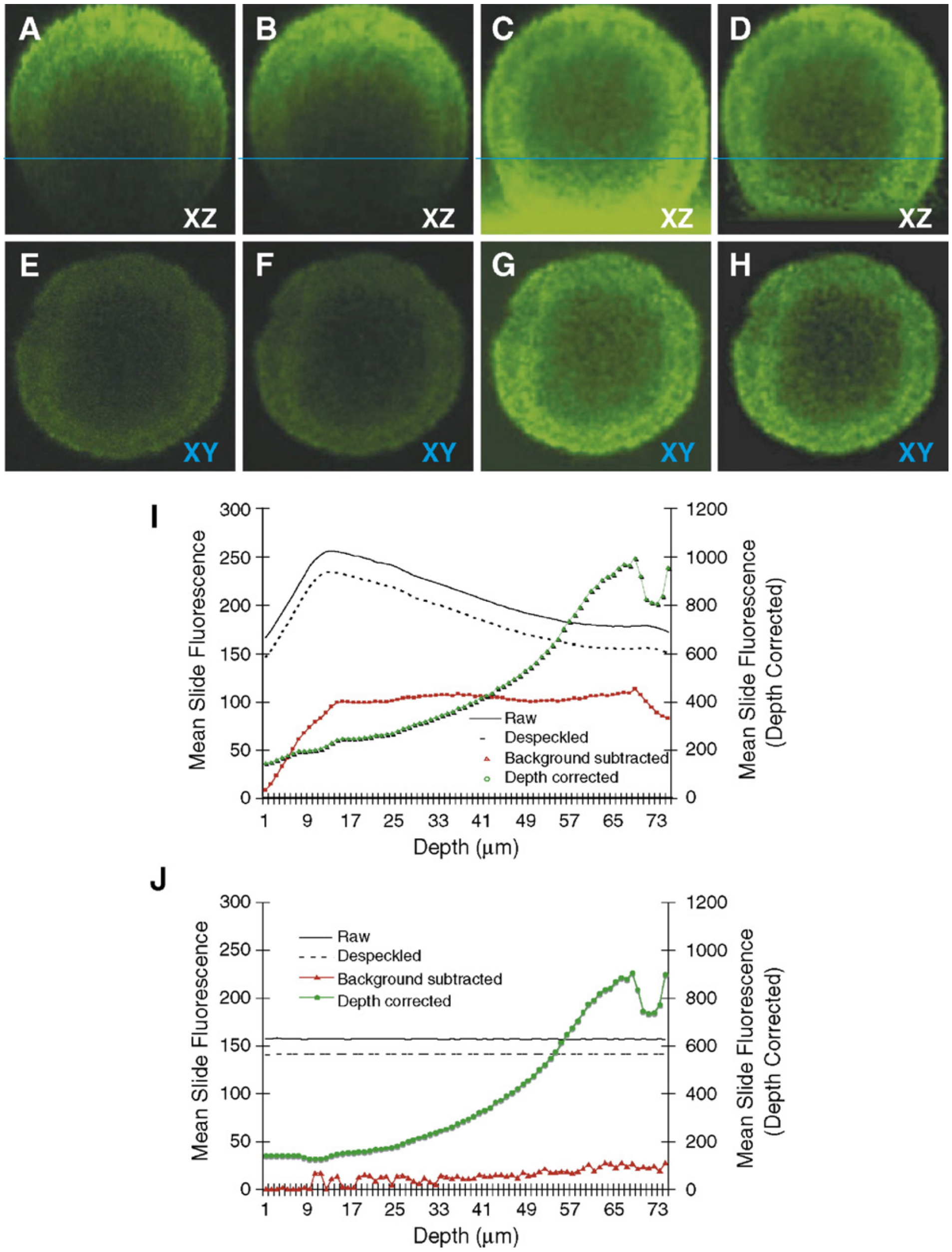


Figure. 3.1. Depth and cross-sectional profiling of eggs injected with GFP protein. (*A-D*) XZ and (*E-H*) XY cross-sections of injected eggs after consecutive stages of image-processing. The XZ cross-sections are oriented such that the top part of the image is facing the confocal lens. A blue line represents the slice in the XZ profile from which the XY image was taken (45.6 μm deep). (*A, E*) Unprocessed stack, showing depth-dependent signal decay. (*B, F*) Image stack after applying the median-filter (despeckling) algorithm. (*C, G*) Image stack after depth correction showing increased background signal. (*D, H*) background-subtracted stack showing reduced background and uniform cellular fluorescence along the Z-axis. (*I*) Depth profile showing mean slice intensity in an unprocessed image stack and after each subsequent filtering procedure was applied. (*J*) Depth profile of the dark background in the image stack. Dark background profile

was obtained by measuring fluorescence in a region of the field that did not overlap with the egg. The section dimensions were 230 microns in length and width and 1.2 microns thick.

the green channel stack is then normalized by dividing intensities for each slice by their corresponding red depth profiles and multiplying by the maximum green mean intensity (i.e., the mean green intensity for the most shallow section in the stack, following background correction). The correction algorithm is applied indiscriminately to all pixels in a stack, and so one consequence is that it inflates the background fluorescence in deeper images of the embryo (Fig. 3.1*J*). To deal with this, background image fluorescence (dark noise) throughout the stack is estimated by sampling pixel intensities in a square region of each image that does not overlap the sample. Processing is complete once this estimate of dark noise is subtracted from each image. The application of this method adequately corrects for depth-dependent loss of signal, but we have found that it leads to a 10-20% loss of fluorescence in the lowest quarter (deepest 20 microns) of the imaged embryo. We improved upon this method by implementing a Gaussian blurring algorithm that compensates for loss of signal in deep slices. The empirical application of a Gaussian blur to the fluorescence data in each section has the effect of interpolating missing data in the deeper slices of the image stack, at a cost of minor loss of precision in measurement. Figure 3.2 compares red channel signal in a 24 h embryo without Gaussian blurring and with blurring using a 1 or 2 pixel radius filter. The Gaussian algorithm redistributes the pixel intensity of any given spot over its neighbors along a Gaussian curve that extends either 1 or 2 pixels away in all directions. Using a Gaussian filter with radius of 1 pixel, the number of thresholded voxels in an image stack correlates well with that predicted for a spherically-shaped embryo with an outer diameter of 88 microns and an inner diameter (blastocoel) of 63 microns. In this experiment, if the filter is not applied, there is a 20% loss of signal in the deepest slices of the stack. A Gaussian filter of radius two pixels overcompensates in deep sections, producing a slight fluorescence peak at 80 microns, so a filter of radius one was used for all future corrections. The noticeable peak in the number of thresholded pixels between 25 and 55 microns in Fig 3.2 is attributed to the ingressing mesenchyme cells present in 24 h blastula-stage embryos.

The data in Figures 3.1 and 3.2 demonstrate both the magnitude of the depth problem for quantitative imaging of reporter gene expression, and the effectiveness of the solution that has emerged from this work. This enables equivalent collection of relative signal from any location in the embryo. The next objective was to convert relative signal to absolute numbers of GFP protein molecules.

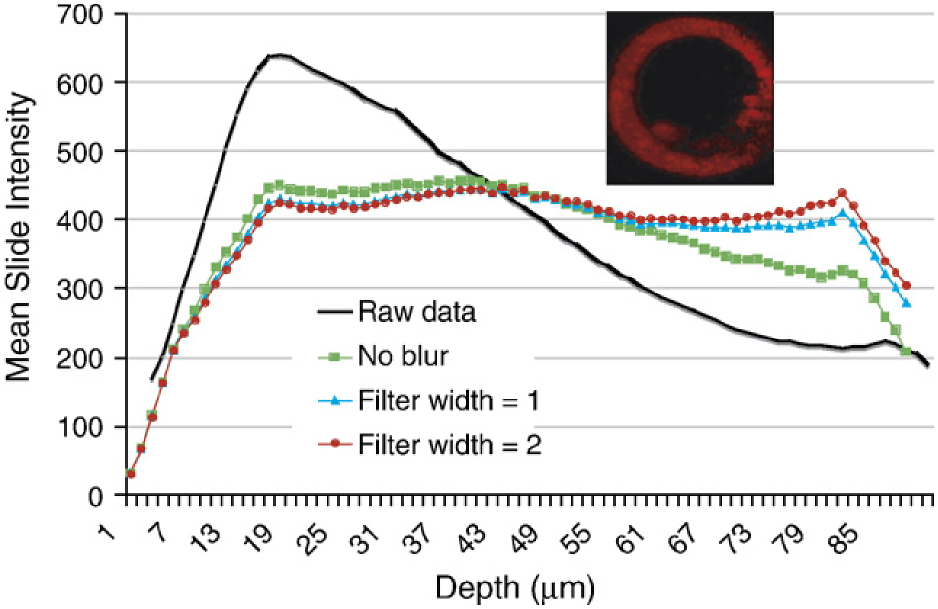


Figure. 3.2. Application of Gaussian filters to compensate for loss of signal in deep image stacks. (*A*) XZ cross-section of a 24 hpf embryo (red channel) processed using the depth-correction algorithms described in text and a Gaussian filter set to a value of 1 pixel. (*B*) Comparison of mean slide intensities with varying degrees of Gaussian blur. Mean slice intensities are shown for raw (despeckled) image stack (yellow), and depth-corrected image stacks with no Gaussian filter (teal) and with Gaussian filters of pixel radius 1 (purple) and 2 (dark brown-red).

From fluorescence intensity to GFP mass.

Pixel intensity is a unitless value (range 0 to 65535). To translate intensity into the biologically relevant number of GFP molecules produced by the reporter genes in a given cell, we employed a direct standard. Fertilized eggs were injected with a known concentration of purified GFP protein in a measured volume. The injection volume, roughly 10 picoliters, was derived by measuring the diameter of the injection bubble using an ocular micrometer. The injection bubble diffuses rapidly upon entering the egg, preventing an accurate measurement of bubble dimensions. However, the addition to the injection buffer of 12.5% glycerol partially obviates this difficulty and aids in boundary detection, as it increases the viscosity of the injection solution and retards its rate of diffusion in the egg cytoplasm. These eggs were then imaged by CLSM to produce a stack of about 60 images, each representing a 1.2 µm-thick section of the embryo. Table 1 shows the corrected fluorescence intensity in six injected eggs after processing the image stack as described above. The total fluorescence (Ft) was measured as the sum of all processed voxel intensities within the stack, and the GFP fluorescence (G) was calculated by subtracting from Ft the background cell fluorescence (Fb) obtained from a mock injection (G = Ft–Fb). The mean total fluorescence was $6.9x10^8$; the background fluorescence was $1.2x10^8$; and mean GFP fluorescence in the injected eggs was thus measured to be $5.8x10^8$ (absolute units of intensity in a series of 16-bit tiff images). The conversion factor of 0.18 molecules/16-bit intensity unit was then obtained by dividing the mean number of GFP molecules injected ($1.07x10^8$) by the total GFP fluorescence intensity ($5.8x10^8$).

Table 3.1. Measurements of fluorescence in fertilized eggs injected with purified rEGFP protein

Egg	Corrected fluorescence
1	1.9E+08
2	1.0E+09
3	7.4E+08
4	5.9E+08
5	5.3E+08
6	3.9E+08
Mean	5.8E+08

The fluorescence generated by the injected protein (BD Biosciences, cat#632439) has been corrected by subtraction of background. The mean injection volume was 10 pl. and the mean number of GFP molecules injected was 1.07×10^8.

Measurement of GFP expression in individual cells expressing a Tbrain-GFP BAC reporter

To apply these methods to an actual experimental situation, we utilized a GFP expression construct active specifically in the skeletogenic cells of the sea urchin embryo. Skeletogenesis occurs after gastrulation, and is executed within the blastocoel by a specific lineage of cells, 32 in number at this stage. These cells align themselves in a bilaterally symmetrical pattern on the ectodermal wall of the blastocoel, evidently directed by signals expressed in the ectoderm (Hodor and Ettensohn, 1998). They form a syncytial cable-like structure within which the skeletal biomineral-protein complex is secreted. This is directly relevant to our present concerns because in later embryos (after about 40 h) the syncytium permits the GFP product of an expression vector incorporated in some of the mesenchyme cells to diffuse to all other mesenchyme cells, facilitating visualization of the whole skeletogenic structure. Incorporation of injected *cis*-regulatory constructs in sea urchin eggs is a mosaic process, and typically in these experiments, for example, one-fourth of the 32 cells may contain the exogenous DNA. The skeletogenic cell lineage descends from four specific 5th cleavage blastomeres (the “large micromeres”), and their fate is specified very early in development by a known set of regulatory gene interactions (Oliveri and Davidson, 2004). Among the regulatory genes expressed early in this cell lineage is *tbrain*, which encodes a T-box family transcription factor (Croce et al., 2001; Davidson et al., 2002; Oliveri and Davidson, 2004). Transcription of the *tbrain* gene begins in late cleavage (12 hpf) and persists throughout PMC ingression and skeletogenesis (Hodor and Ettensohn, 1998; Croce et al., 2001).

The expression construct we used in the following experiments consisted of a BAC containing the *tbrain* gene, into the first exon of which the GFP coding sequence had been inserted by reciprocal recombination (Yu et al., 2000). In embryos grown from eggs injected with this construct the GFP reporter is expressed with perfect fidelity exclusively in the skeletogenic cell lineage, exactly as is the endogenous *tbrain* gene (data not shown).

Fertilized eggs injected with the Tbrain BAC construct were grown to the 24 h mesenchyme blastula stage, when the skeletogenic cells have just ingressed into the interior of the embryo, where they lie in an easily recognized pile at the vegetal end of the blastocoel. To quantitate the GFP expression on a per cell basis, processed CLSM images were segmented to isolate the signals from individual cells. Individual cell fluorescence was measured by defining a cylindrical volume of interest: the diameter was defined by the cell's major axis and the length was defined by the number of sections in which the cell was captured. Total fluorescence was obtained by summing the intensity of all thresholded voxels within this cylinder. For these measurements a threshold of 150 intensity units was set empirically by measuring mean background fluorescence in uninjected embryos.

In Figure. 3.3a is shown an optical slice of a 24 h embryo in which four mesenchyme cells can be seen expressing the Tbrain-GFP reporter (red color). Inactive mesenchyme cells can also be seen within the blastocoel. The calculated GFP intensity for each slice of a fluorescent mesenchyme cell and a nonfluorescent cell are shown in Fig. 3.3b. The summed voxel intensity of the fluorescent cell is $8.9\text{x}10^6$ and the summed voxel intensity of a nonfluorescent cell of similar dimensions is $1.8\text{x}10^6$, so after correction for background the intensity of reporter expression is 7.1x10^6, about four times background. The calculated number of GFP molecules in the fluorescent cell is then 2.8x10^5. The background autofluoresence sets the practical sensitivity of the method. From these measurements the number of GFP molecules required to equal the autofluorescence per cell is about 7x10^4.

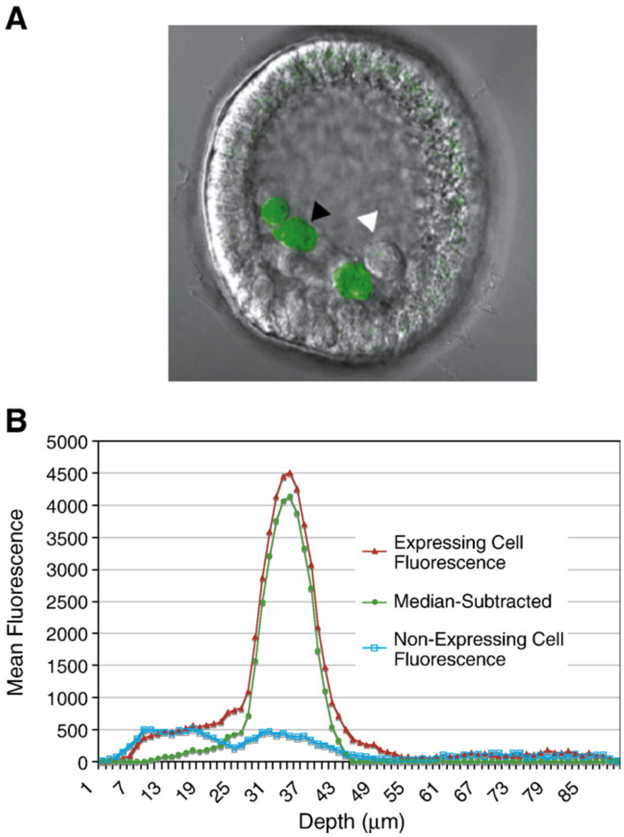


Figure. 3.3. Depth profiles of GFP for fluorescent and non-fluorescent skeletogenic mesenchyme cells. (A) CLSM image of a 24 hpf embryo injected with a GFP-reporter under control of the SpTbrain *cis*-regulatory system. Black arrow points to the fluorescent cell measured in (B) and white arrow points to a nonfluorescent cell. (B) Mean fluorescence intensity of a GFP-expressing cell (yellow triangle) and a nonfluorescent cell (blue hyphens). The median intensity of nonfluorescent cells provides a consistent measure of background cell fluorescence, and can be subtracted from the total fluorescence of the cell (red circles).

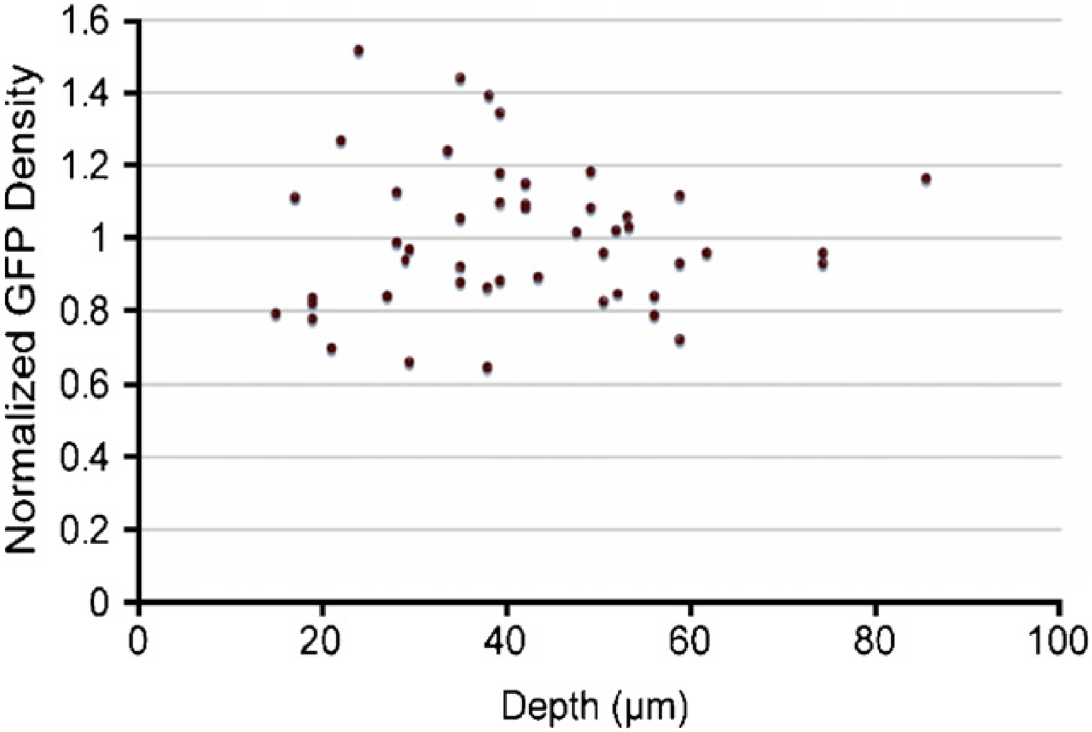


Figure. 3.4. Normalized GFP density vs depth. GFP density was normalized by the mean GFP density of all fluorescent skeletogenic cells in a given embryo.

Quantitation of Tbrain-GFP expression in individual syncytial skeletogenic cells.

Embryos bearing the Tbrain-GFP expression construct were grown to a late gastrular stage (52 h) and imaged using CLSM. At this stage the skeletogenic cells are arranged in a three dimensional syncytial structure and it is not possible to include them all in any single optical section. Multiple individual cell bodies were imaged from each of several embryos, with the results described in Fig. 3.5*C*. Here is shown the number of GFP molecules per cell, for each of five embryos, calculated by multiplying the summed voxel intensity after correction for background, by use of the conversion constant derived above, i.e., 0.18 GFP molecules per 16-bit intensity unit. Figure 3.5*C* indicates that total GFP mass varies between 7x10^5 and 2.6x10^6 molecules per cell (standard deviation 50%).

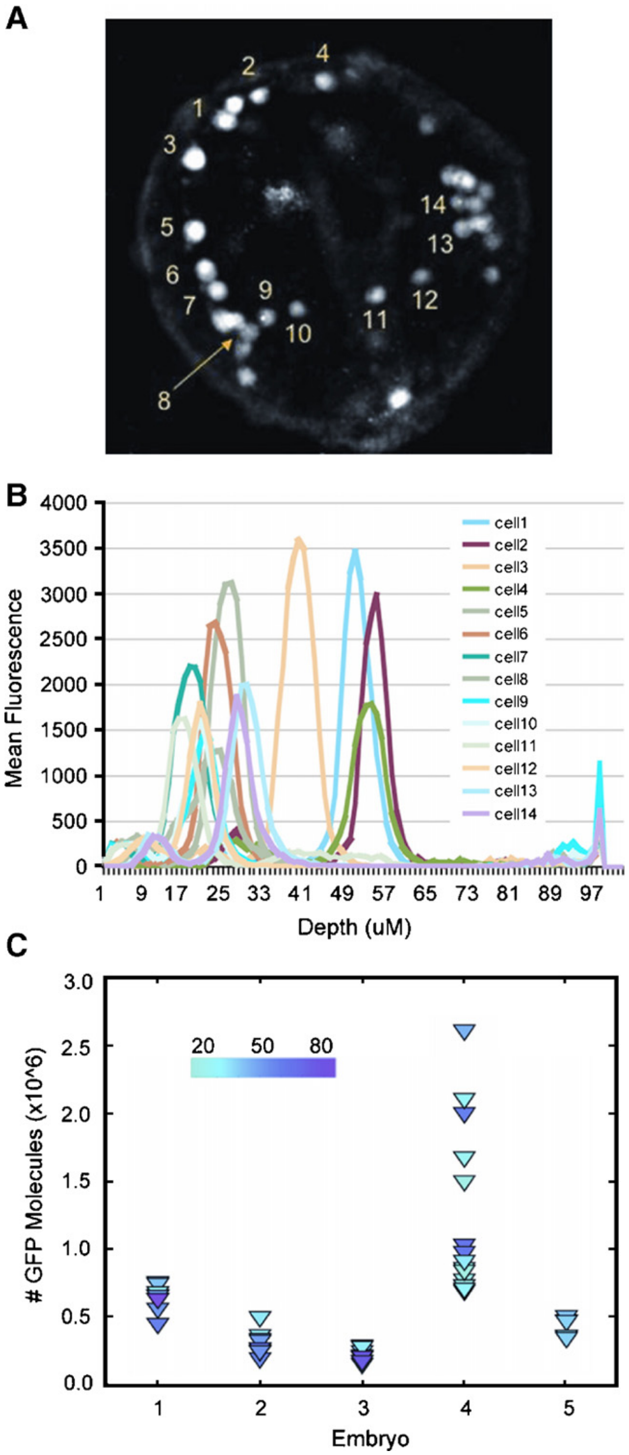


Figure. 3.5. Quantitative estimation of GFP content in individual skeletogenic cells expressing a Tbrain-GFP BAC knock-in. Data are from Embryo 4 of Fig. 3.5c. (A) Z-projection of CLSM image stack taken at late gastrula stage. The measured cells are numbered 1-14. (B) Fluorescence profiles (as in Fig. 3.3) for all 14 cells. (C) Calculated GFP contents of individual cells and colored by depth (μm) for embryos injected with

Tbrain-GFP BAC expression construct (embryos 1-4) or linearized Tbrain-GFP plasmid expression construct (embryo 5).

We asked whether the depth correction algorithm effectively removed any correlation between signal intensity and depth of imaging in this data set. Due to heterogeneity in the copy number of integrated reporters, the GFP density per mesenchyme cell between different embryos will always be different, though each active cell in a given embryo will contain the same number of exogenous DNA copies. To allow for direct comparison of GFP densities between embryos, the measured GFP density per cell was normalized by the average GFP density of all fluorescent mesenchyme cells measured in a given embryo. Figure 3.4 shows that there is no correlation between depth and calculated GFP activity, and deeper cells show the same range of GFP fluorescence per unit cell volume as do cells closer to the surface. Thus, for example, the normalized GFP density of the 10 most shallow cells, which occupy depths between 15 and 28 μm, have a mean of normalized fluorescence density of 0.96 +/- 0.25 whereas that of the deepest 10 cells, which occupy depths between 53 to 85 μm, have mean densities of 0.95 +/- 0.14.

Are the calculated numbers of GFP molecules consistent with the amount of GFP mRNA generated by the expression constructs in these embryos? GFP mRNA abundance was measured by quantitative PCR (QPCR) in batches of 100 embryos injected with a plasmid Tbrain-GFP construct including 3.5 kb of *tbrain cis*-regulatory sequence, which faithfully reproduces *tbrain* expression in skeletogenic mesenchyme (data not shown). This construct was used rather than the BAC knock-in in order to increase the molar quantity of incorporated exogenous expression constructs and hence the output of GFP mRNA. A representative sample of the transgenic embryos was then imaged by CLSM, and GFP fluorescence per mesenchyme cell body was computed (Figure 3.5c, Embryo 5). The QPCR data yield an abundance of 4.75x10^4 molecules of GFP mRNA per embryo. Given that there are 32 syncytial skeletogenic mesenchyme cells at midgastrula stage and that these distribute their mRNA evenly, it can be assumed that on average there are ~1400 molecules per cell. The endogenous expression of *tbrain* peaks at roughly 5000 copies per embryo at 24 hpf and falls to 3000 copies per embryo at mid-gastrula stage. Our measurements thus indicate a GFP abundance 8-14 times that of endogenous *tbrain* mRNA in each cell, which is quite reasonable considering that multiple copies of the reporter gene are incorporated.

The mean GFP protein abundance per cell in these embryos, as calculated from the GFP fluorescence, was 1.2x10^6 molecules. At the sea urchin embryo translation rate of two protein molecules per mRNA-minute (Davidson, 1986), then, it would require only 7 h to produce the measured 1.2 million molecules of GFP protein from 1400 messages. This is easily within the range expected, since GFP protein is very stable; and since endogenous expression is already up at 24 h, while these measurements were made a whole day later, at 46-52 h.

In Fig. 3.5 we demonstrate the use of these methods to resolve the basic problem adduced in the introduction, that is, obtaining both spatial and quantitative measurements of reporter gene expression on the same embryo. The projection of optical CLSM sections of different depths shown in Fig. 3.5a displays multiple syncytial skeletogenic cells expressing the BAC-GFP construct. Fluorescence profiles for 14 individual cells are shown in Figure 3.5b. The calculated GFP molecules in each individual cell are shown in Figure 3.5c, plotted in respect to depth of the cell in the imaged embryo. The cells all fall within a factor of two of their mean calculated GFP content, but there is no correlation between GFP content and their depth in the embryo. These small differences may be real; for example diffusion of GFP among all the cells of the syncytium may be limited both in time and space, resulting in differences in GFP content between cells whose genomes do or do not contain copies of the reporter DNA construct.

DISCUSSION

It is now possible to combine the beautiful high resolution imaging afforded by CLSM with quantitative measurement of reporter gene output in a living embryo. This opens an entirely new range of opportunities for cell-by-cell *cis*-regulatory analysis, in which both spatial and quantitative expression functions are included. The requirement for quantitative assessment of reporter gene output is obvious: many of the target sites in *cis*-regulatory modules control various aspects of the amplitude of gene expression (e.g., Yuh et al., 1998; Yuh et al., 2001), while others determine whether the gene will be expressed in a given place in the embryo. The overall organization of these genomic control systems cannot be understood unless both kinds of function [and others as well, (Istrail and Davidson, 2005)] are taken into account. In *cis*-regulatory analysis the canonical approach is to measure the functional effects of mutations of given target sites in an expression construct, and now, at least in sea urchin embryos, both quantitative and spatial effects can be dealt with by the same measurement protocol, in the same embryo.

This approach can be easily extended to model systems, such as ciona, starfish and zebrafish, for which gene transfer is an important experimental methodology and for whose physical dimensions and optical transparency are similar to that of the sea urchin. There are no other physical limitations barring the extension of this approach to these systems. The measurement of kinetics of reporter gene expression in a given embryo over time, is an objective of current efforts and will be greatly facilitated by recent innovations in rapid confocal microscopy being developed in the Fraser lab. Many other new opportunities now present themselves, such as the use of multiple color reporters to compare in the same cells the behavior of different mutations of a given *cis*-regulatory module; or the study of signaling perturbations on the expression of given genes, as these often affect both spatial and quantitative output. In sum, we have found a way to escape the general exclusion between spatial and quantitative measurement of *cis*-regulatory activity for at least one developmental system. Variants of this method should be widely applicable to the many other systems in which the same problem obtains.

ACKNOWLEDGEMENTS

This work was supported by NIH grant HD-37105 and Department of Energy grant DE-FG02-03ER63584. Confocal images were taken at the Biological Imaging Center in the Beckman Institute at the California Institute of Technology.

REFERENCES

Abramoff, M. D., Magalhaes, P. J., and Ram, S. J. (2004). Image processing with ImageJ. Biophotonics International 11, 36-42.

Arnone, M. I., Dmochowski, I. J., and Gache, C. (2004). Using reporter genes to study cis-regulatory elements. Methods Cell Biol. 74, 621-52.

Chalfie, M., Tu, Y., Euskirchen, G., Ward, W. W., and Prasher, D. C. (1994). Green fluorescent protein as a marker for gene expression. Science 263, 802-5.

Croce, J., Lhomond, G., Lozano, J. C., and Gache, C. (2001). ske-T, a T-box gene expressed in the skeletogenic mesenchyme lineage of the sea urchin embryo. Mech. Dev. 107, 159-62.

Davidson, E. H. (1986). Gene Activity in Early Development, third edition. Academic Press, Orlando, FL, pp.126-193.

Davidson, E. H. et al. (2002). A genomic regulatory network for development. Science 295, 1669-78.

Dmochowski, I. J., Dmochowski, J. E., Oliveri, P., Davidson, E. H., and Fraser, S. E. (2002). Quantitative imaging of cis-regulatory reporters in living embryos. Proc. Natl. Acad. Sci. U. S. A. 99, 12895-900.

Hodor, P. G., and Ettensohn, C. A. (1998). The dynamics and regulation of mesenchymal cell fusion in the sea urchin embryo. Dev. Biol. 199, 111-24.

Istrail, S., and Davidson, E. H. (2005). Logic functions of the genomic cis-regulatory code. Proc. Natl. Acad. Sci. U. S. A. 102, 4954-9.

Lee, P. Y., and Davidson, E. H. (2004). Expression of Spgatae, the Strongylocentrotus purpuratus ortholog of vertebrate GATA4/5/6 factors. Gene Expr. Patterns 5, 161-5.

McMahon, A. P., Flytzanis, C. N., Hough-Evans, B. R., Katula, K. S., Britten, R. J., and Davidson, E. H. (1985). Introduction of cloned DNA into sea urchin egg cytoplasm: replication and persistence during embryogenesis. Dev. Biol**.** 108, 420-30.

Oliveri, P., and Davidson, E. H. (2004). Gene regulatory network controlling embryonic specification in the sea urchin. Curr. Opin. Genet. Dev. 14, 351-60.

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, 438-51.

Wang, D. G., Britten, R. J., and Davidson, E. H. (1995). Maternal and embryonic provenance of a sea urchin embryo transcription factor, SpZ12-1. Mol. Mar. Biol. Biotechnol. 4, 148-53.

Yu, D., Ellis, H. M., Lee, E. C., Jenkins, N. A., Copeland, N. G., and Court, D. L. (2000). An efficient recombination system for chromosome engineering in Escherichia coli. Proc. Natl. Acad. Sci. USA. 97, 5978-83.

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, 1896-902.

Yuh, C. H., Bolouri, H., and Davidson, E. H. (2001). Cis-regulatory logic in the endo16 gene: switching from a specification to a differentiation mode of control. Development 128, 617-29.

Yuh, C. H., Dorman, E. R., and Davidson, E. H. (2005). Brn1/2/4, the predicted midgut regulator of the endo16 gene of the sea urchin embryo. Dev. Biol. 281, 286-98.

Yuh, C. H., Dorman, E. R., Howard, M. L., and Davidson, E. H. (2004). An otx cis-regulatory module: a key node in the sea urchin endomesoderm gene regulatory network. Dev. Biol. 269, 536-51.

CONCLUSIONS

In this work, I pursued two methods for validating gene regulatory network structure. First, I rewired the gene regulatory network for specification of skeleton by appending to it the subcircuit responsible for driving pigment cell specification. The outcome of this modification was to divert the skeletogenic fate towards pigment cell fate. This change both confirmed aspects of existing understanding of the pigment cell specification network and also revealed a novel cross-repressive activity of the pigment cell fate against skeletogenic fate. Second, I performed a cis-regulatory analysis on the alx1 gene and showed it is directly regulated by the double-negative gate of *pmar1* and *hesc* and that its expression is initiated by ets1. I also identified the regulatory logic responsible for its dynamic early expression kinetics was controlled by the dual role of alx1 protein as both an autoactivator and delayed autorepressor.

REWIRING DEVELOPMENTAL GENE REGULATORY NETWORKS

That the GRN rewiring performed in Chapter 1 was capable of completely reprogramming specification and differentiation has reinforced some notions about the importance of network topology and logic in directing development. In summary, the expression of *gcm* was brought under control of the double-negative gate and ets1 activation by borrowing the cis-regulatory architecture of *tbrain*. This regulatory rewiring caused *gcm* to be expressed at a very early time in the specification of skeletogenic precursors, essentially prior to commitment of this lineage to a skeletal fate and in synchrony with subcircuitry for the initiation of skeletogenesis. As a result, *gcm* initiated a sequence of regulatory events that led to the stabilization of the pigment cell state and ultimately to deployment of pigment differentiation battery and concomitant loss of expression of the skeletal differentiation battery. The successful cell fate reprogramming was possible due to a combination of a number of factors, which I will described in greater detail in the following section:

A. Repression of critical early regulators of skeletogenesis
B. *Gcm* expression was driven to high levels in the SM lineage.
C. Synthetic expression occurred prior to lockdown of SM specification state
D. The initiation of endogenous *gcm* was removed from control of the delta-notch system
E. Synthetic expression triggered the regulatory subcircuit required to lock down pigment cell fate.

Repression of critical early regulators of skeletogenesis

The transcription factors *alx1, ets1* and *tbrain* are responsible for initiating skeletogenesis. *Alx1* and *ets1* are direct targets of the differentiation battery genes expressed during skeletogenesis and also of the molecular machinery responsible for SM cell epithelial to mesenchyme transition and subsequent migration and syncytium. *Tbrain,* is unnecessary for EMT transition, however is required along with *ets1* for triggering specification state lockdown circuitry involving *erg, hex* and *tgif.* The synthetic expression of *gcm* in the SM lineage was capable of downregulating the transcription of all three early initiators of skeletogenesis: *alx1, ets1,* and *tbrain* causing a delay in ingression and a failure of ingressed SM cells to arrange in bilateral clusters and fuse to form syncytium. The lowered levels of *tbrain* and *ets1* may also have contributed to the failure to lockdown the differentiation program. While the levels of *hex, erg* and *tgif* were not directly measured here, it is clear that the expression of differentiation battery genes like msp130L are severely effected by synthetic expression of gcm. Given that differentiation battery genes are generally wired in parallel "OR" configuration with multiple positive inputs, the severe loss of msp130 is an indirect indication that all or nearly all of the positive SM regulators of differentiation are downregulated in the context of synthetic *gcm* expression. For *msp130L* these factors include, *hex, erg, ets1,* and *alx1.*

Gcm expression was driven to high levels in the SM lineage

The *tbrain* gene is zygotically expressed beginning at 8-10 hpf and reaches about 2000 copies per embryo at its peak expression point at around 20 hpf. Given that there are 8 SM cells by 20hpf, this corresponds to roughly 250 copies tbrain per cell, or 125 copies per *tbrain* allele. It was calculated in Chapter 1 that the Tbrain-GCM BAC was capable of expressing synthetic *gcm* at roughly 40 copies per construct. By accounting for the level of mosaicism and numbers of integrated BAC constructs, I was able to calculate that the expression of synthetic *gcm* per cell was in the range of at least 200cp/cell at 11.5hpf (relatively soon after initiation) and over 800 copies/cell by 20 hpf. *Gcm* MOE experiments independently showed that a level of 280 copies *gcm* mRNA per cell was sufficient to reduce Alx1 expression at 11.5 hpf by 50% and to drive strong expression of the pigment cell marker *pks*. These results indicate that the *tbrain* cis-regulatory architecture provided a sufficient dosage of *gcm* to permit its cross-repressive functions on skeletogenesis and to permit the deployment of the pigment cell differentiation program. Interestingly, the choice to use *tbrain* cis-regulatory architecture over *alx1* was a fortuitous one in that *tbrain* zygotic expression is initially influenced by *ets1* alone whereas, as was discovered in Chapter 2, the high levels of *alx1* transcription requires the *alx1* input itself. *Gcm* driven by the *alx1* cis-regulatory architecture would therefore likely not have reached sufficient levels to permit respecification.

Synthetic expression occurred prior to lockdown of SM specification state

As was described in the Introduction, the genes *hex, erg* and *tgif* form an critical 3-gene subnetwork that is important for stabilizing the skeletogenic regulatory state. Their expression is triggered by *ets1* and *tbrain* at 15-20 hpf, but soon becomes independent of those inputs through extensive positive cross-regulatory wiring. In short, *erg* and *hex* engage in a cross-activating loop and they both activate *tgif*, whose role is as a positive input in late *alx1* expression and also as a second maintenance input for *hex*. Given the potent cross-regulatory wiring of this state-stabilizing subnetwork, it seems unlikely that synthetic *gcm* expression would have diverted the skeletogenesic program had it not occurred prior to expression of *erg* and *hex* at 15-17hpf. In fact, if as a hypothetical example the *hex* cis-regulatory architecture were used to drive *gcm,* it might be possible to activate both differentiation programs within the same cell to produce a hybrid skeleton with additional pigment cell morphologies.

Initiation of endogenous gcm was removed from control of the delta-notch system

Gcm expression is regulated by an early and late cis-regulatory module, as was described in the Introduction. Initially, delta-notch signaling by adjacent SM precursors interact directly on the gcm cis-regulatory architecture through Suppressor of Hairless binding sites. These sites act to restrict *gcm* expression solely to the NSM precursors. By bringing *gcm* under the control of the *tbrain* cis-regulatory wiring, the dependence on notch for early activation and restriction to the NSM lineage was removed. This allowed synthetic *gcm* to be initiated in the SM lineage, although endogenous *gcm* was still repressed in these cells, as notch signaling is inactive here. *GCM* MOE experiments showed however that ectopic expression of gcm is capable of activating endogenous gcm. This form of activation likely occurs through the late *gcm* module, which is believed to contain *gcm* binding sites that permit autoactivation (A. Ransick, unpublished results). The activation of endogenous *gcm* was an important step in the successful respecification of SM to the pigment cell lineage because it freed *gcm* from control by the *tbrain* cis-regulatory architecture and by initiation through *ets1*. The rewiring therefore created a scenario where the *tbrain* cis-regulatory architecture provided primarily an "ignition" switch for activating *gcm* but which later became unnecessary for maintenance of gcm levels.

Synthetic expression triggered the regulatory subcircuit required to lock down pigment cell fate

The synthetic expression of *gcm* was able to activate the endogenous gcm cis-regulatory architecture itself through the late module, and by the activation of other factors such as *six1/2* which also appear to feed back positively on *gcm* and stabilize its expression. Additionally, experiments by S. Materna implicate *gcm* as a strong co-activator of *gata-e* expression (unpublished). *Gata-e* and *gcm* are both required to activate expression of a number of pigment-cell specific genes. Thus, *gcm* expression is sufficient to activate the subnetwork required for stabilization of the pigment specification state, and also for activation of a pigment cell differentiation program.

The details of the mechanism by which *gcm* represses *alx1* and *ets1* transcription remain to be solved. A simple explanation, given the results from chapters 1 and 2, is that *gcm* directly represses *ets1* transcription and that loss of *ets1* leads to loss of *alx1.* In partial support of this idea a coinjection of a minimal reporter (J-module and alx1 basal promoter) displayed the same 2-fold reduction of expression as the alx1-GFP BAC when coinjected with the Tbrain-GCM BAC. This result is consistent with the proposed mechanism, but conclusive evidence can only come from a cis-regulatory analysis of the ets1. A second alternative is that *gcm* independently represses both *alx1* and *ets1.* If this were the case, it would be expected that gcm binding sites could be found within the J-module. A comparison of *gcm* DNA-binding domains from orthologs in *S.purpuratus* and *D.melongaster* shows they are highly conserved (Cohen et al., 2003) and the canonical bining sequence for *gcm* is $^{G}/_{A}{}^{C}/_{T}$CCGCAT (Akiyama et al., 1996; Miller et al., 1998). Also, functional *gcm* binding sites of the form $^{G}/_{A}$CCCGCAT have been identified within the *Sp-gcm* and *Sp-pks* loci. A search for these sequences within the J-module provided no strong matches, suggesting *gcm* is not a direct regulator of *alx1*.

CIS-REGULATION OF SKELETOGENIC REGULATORY FACTORS

In Chapter 2, I identified the key regulators responsible for initiation and maintenance of early *alx1* expression in the SM lineage. A final mystery regarding *alx1* regulation is why this gene is not expressed in NSM cells after hesc clearance at 24 hpf despite the presence of strong ets1 expression in these cells. Initially it was suspected that tbrain cis-regulation might offer a clue because both genes are driven by an ets1 input and both are initially restricted to the skeletogenic micromeres through a promoter-proximal module containing functional Hesc binding sites. However, the *tbr* cis-regulatory architecture capable of NSM expression is repressed there through repression by the Erg transcription factor. At the time when *hesc* clears from the NSM lineage, *erg* and *ets1/2* are coexpressed. In this condition, *erg* blocks *tbrain* activation by outcompeting *ets1/2* for binding sites in the activator module. As might be expected then, morpholino knockdown of *erg* leads to ectopic NSM expression of a Tbrain:GFP BAC. The same behavior, however, was not observed for alx1. When coinjected with *erg* MASO, the alx1 GFP BAC does not show additional expression in NSM lineages. The late expression pattern of *alx1* was not the focus of this study but will certainly reveal the regulatory inputs, both the activation and repression systems, responsible for ensuring accurate expression.

Although both *tbr* and *alx1* are initiated by the same input, *ets1,* the expression of *alx1* peaks a full 7 hours before *tbr*. This fast rise has been shown in this work to be caused by positive autoregulatory feedback by the Alx1 protein. A second question then is what is the developmental significance for this early peak. Among the regulatory factors that are zygotically expressed downstream of the double negative gate, both *alx1* and *ets1* but not *tbr* are required for the precocious ingression of SM precursors beginning at 22-24 hpf. Between these two *alx1* is directly upstream of the ingression circuitry since *ets1* is a direct input to *alx1,* and since *alx1* MOE alone is sufficient to induce EMT in the entire embryo. It is possible then that early and fast peak of *alx1* expression at 11.5 hpf is a requirement for ensuring the relatively early timing of ingression of these cells, and that a delay in high *alx1* levels would correspondingly delay EMT. Some evidence for this model exists from experiments performed by Smith and Davidson (Smith and Davidson, 2009) and described in the Introduction, which reveal a blimp1-dependent failsafe mechanism for ensuring clearance of *hesc* in the large micromeres. Briefly, both *blimp1b* and *pmar1* are turned on in 4th cleavage micromeres by localized maternal inputs, and *pmar1* immediately represses *hesc* transcription whereas *blimp1b* initially acts as an activator of the wnt8 gene to induce wnt signaling, but later on directly represses both its own expression and that of *hesc*. The mechanism of this delayed repression is not well understood, however the timing was approximated to be roughly 6 hours after *pmar1* expression, or at roughly 11-13 hpf. Correspondingly, knockdown of pmar1 through MASO injection only postponed ingression from 22-24 hpf to 30hpf and did not block skeletogenesis. That the length of the ingression delay roughly equals the length of delay of *hesc* clearance suggests that a minimum amount of time, 16-18 hours, is required between the activation of the double negative gate and EMT. Since *alx1* is both necessary and sufficient for EMT, as shown in this work, and it is expressed as a direct consequence of *hesc* clearance, then its expression is the determining factor for timing ingression of SM cells in normal development.

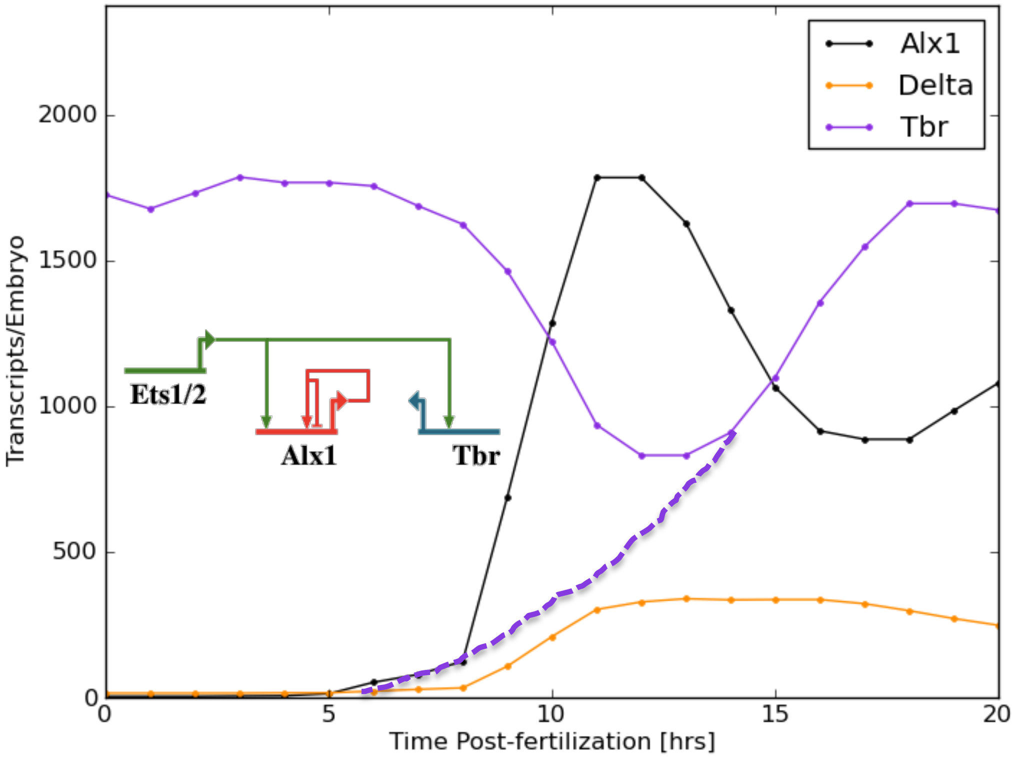


Figure 4.1. Kinetic timecourse of *alx1, tbr and delta* during the first 20 hours of development (Materna et al., 2010). *Tbr* is expressed maternally, so its initial zygotic expression pattern is estimated from its kinetics and from the cis-regulatory analysis of *tbr* (Wahl et al., 2009) (dashed purple line). Inset: cis-regulatory organization of *ets1/2, alx1* and *tbr* in the large micromeres.

Another importance for strong and early *alx1* was indicated by experiments performed by P. Oliveri and Q. Tu (Oliveri et al., 2008) which showed a cross-repressive role for *alx1* against NSM cell fates. The authors showed that *alx1* knockdown led to ectopic expression of *gcm* in the skeletogenic mesenchyme at late blastula stage. It is therefore possible that the SM precursors retain some potential to develop into NSM lineages, and that early *alx1* expression is therefore required to lock out these alternate regulatory fates in addition to promoting skeletogenesis. The mechanism for expression of *gcm* in SM cells in this perturbation is unclear, since normal initiation of gcm in NSM precursors is downstream of delta/notch signaling and from what is understood of notch in the related sea urchin *Lytechinus variegatus* the SM lineage does not express notch protein after 12 hpf ((Sherwood and McClay, 2001). Nevertheless, it is clear from experiments in Chapter 1 that the forced expression of *gcm* in the SM lineage acts to the detriment of skeletogenic differentiation, and therefore the repression of *gcm* here is critical for proper development.

REFERENCES

Akiyama, Y., Hosoya, T., Poole, A.M., Hotta, Y., 1996. The gcm-motif: a novel DNA-binding motif conserved in Drosophila and mammals. Proc Natl Acad Sci U S A 93, 14912-14916.

Cohen, S.X., Moulin, M., Hashemolhosseini, S., Kilian, K., Wegner, M., Muller, C.W., 2003. Structure of the GCM domain-DNA complex: a DNA-binding domain with a novel fold and mode of target site recognition. EMBO J 22, 1835-1845.

Materna, S.C., Nam, J., 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, 177-184.

Miller, A.A., Bernardoni, R., Giangrande, A., 1998. Positive autoregulation of the glial promoting factor glide/gcm. EMBO J 17, 6316-6326.

Oliveri, P., Tu, Q., Davidson, E.H., 2008. Global regulatory logic for specification of an embryonic cell lineage. Proc Natl Acad Sci USA 105, 5955-5962.

Sherwood, D.R., McClay, D.R., 2001. LvNotch signaling plays a dual role in regulating the position of the ectoderm-endoderm boundary in the sea urchin embryo. Development 128, 2221-2232.

Smith, J., Davidson, E.H., 2009. Regulative recovery in the sea urchin embryo and the stabilizing role of fail-safe gene network wiring. Proc Natl Acad Sci USA 106, 18291-18296.

Wahl, M.E., Hahn, J., Gora, K., Davidson, E.H., 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, 428-441.

APPENDIX A

SANN: A Web Service for Integrating Phylogenetic Footprinting, Binding Site and Sequence Database Searches for Cis-Regulatory Analysis

Sagar Damle

Division of Biology, California Institute of Technology, Pasadena CA 91106, USA

(In preparation, *Nucleic Acids Research*, Webserver issue)

ABSTRACT

Gene regulatory network analysis has become a powerful method for understanding biological pathways controlling specification and differentiation. A key step in this process is to identify and validate by mutation the function of putative transcription factor binding sites. Two bioinformatics approaches are currently popular for identifying putative regulatory regions and sites: phylogenetic footprinting and sequence scans against transcription-factor binding site databases. We have created an open-access webtool called SANN, Sequence ANNotation tool, (http://vanbeneden.caltech.edu/~sagar/cgi-bin-pub/sannForm.cgi) that combines these two approaches to analyze single or multiple genomic sequences. The webtool is designed for the cis-regulatory biologist for the purpose of identifying, in multiple

genomic sequences, putative modules that contain a common cohort of transcription factor binding sites. Two tools for phylogenetic footprinting are offered: reciprocal BLAST and PASS-mediated alignment. Several binding site databases can be searched against, including JASPAR, Transfac 2.0, and binding site databases for vertebrate homeodomain and non-homeodomain transcription factors from the UNIPROBE database.

INTRODUCTION

The function of cis-regulatory analysis in the context of gene regulatory network (GRN) analysis is to find direct relationship between a transcription factor onto a target gene. Despite other methods for identifying logical relationships between nodes of a GRN, such as through genetic approaches (conditional knockouts), morpholino knockdown or mRNA overexpression, the method of mutation of binding sites in reporter constructs and BAC reporter constructs remains the unequivocal method for authenticating direct regulatory relationships. (Yuh etal., 2004; Lee etal., 2007).

A general pathway for outlining the use of GFP reporter constructs for studies of cis- regulation in eukaryotic systems has been outlined previously (Smith, 2008). The method begins with the mapping or sequencing of a sufficiently large genomic region, often of 100kb or more in size, that ideally contains both the entire gene coding sequence as well as flanking genomic regions of at least 10-20kb in both directions. The next step, identification of putative cis-regulatory DNA sequences can be greatly facilitated by leveraging homologous sequence information to identify conserved stretches of genomic DNA. These pieces are then tested for spatial and temporal activity by their ability, when ligated to a basal promoter and reporter (such as GFP), to drive expression in a domain overlapping that of the regulated gene. Once functional conserved modules have been isolated, putative transcription factor binding sites can be identified computationally by searching against available binding site databases such as Jaspar (http://jaspar.cgb.ki.se/) and Uniprobe (http://thebrain.bwh.harvard.edu/uniprobe/). A combination of sequence conservation at the nucleotide level, the presence of a shared subset of transcription factor binding sites and biological evidence about which regulatory inputs are expressed at the appropriate time and place can be useful to triangulate functional binding site sequences and reduce false positives.

The sources of binding site information are numerous and not consolidated so that they can be searched all at once. At the same time, free binding-site search tools seldom offer methods for simultaneously looking at sequence conservation at the nucleotide level. Here we offer a free web service that combines two kinds of searches critical for the process of cis-regulatory analysis that is fast, easy-to-use and whose output is intuitive for the biologist.

WEB SERVER FEATURES

SANN contains three sections: A sequence-input field, a search options window and plot options window and a button to submit the sequences for annotation. SANN takes as input a single or multiple FASTA-formatted DNA sequences which can be pasted into a field (see Figure A.1) or uploaded to our server. FASTA-formatted DNA sequences can include degenerate nucleotides (N, R, W, Y, etc...) however these nucleotides are masked before sequences are searched against binding site databases. Large input sequences can be refined by setting a sequence subrange. This is particularly useful after phylogenetic footprinting to zoom in to a region of interest.

Without specifying a particular transcription factor binding site search database, clicking “SUBMIT” will perform the first step to cis-regulatory analysis: a reciprocal sequence search using the BLASTN or PASS algorithm. BLASTN is a well-known tool for aligning stretches of homologous sequence (Altschul etal., 1990). Reciprocal blast (or blast 2 sequences) is therefore a useful way to perform phylogenetic footprinting between two sets of orthologous genomic sequence (von Bubnoff etal., 2005). PASS is a tool written for fast assembly of high-throughput sequence information generated by short-read DNA sequencers (like Solexa) (Campagna etal., 2009). It has been used as well for SNP and IN/DEL detection. In this tool, PASS is used to map short reads of a user-defined length of an input sequence onto a target sequence. The position of the read on the input sequence, and the number of hits on the target sequence are recorded. This information can be displayed on a histogram whereby the abscissa is the length of the input sequence and the ordinate is correlated to the number of matches of the short read to the target sequence. A plot of this type is useful for showing very small islands of conservation (as short as 8-10bp in length) that might not normally be visible through reciprocal blast search as well as identifying short stretches of strong homology within broadly conserved regions. Because SANN is designed to analyze genomic sequences, sequences are also searchable against NCBI refseq databases for the purpose of annotating the position of known transcripts. At this time, only mouse and sea urchin refseq genes can be searched, however additional databases will be added upon request.

Once candidate short (1-2kb) regions have been narrowed down computationally and determined to be functional via reporter assay as described above, they can be further searched for the presence of transcription factor binding sites. An array of transcription factor binding site databases and plotting options are exposed to the user. A brief description of the contents of the binding site databases can be found at the tool website, however there are currently two classes of binding site databases: those containing binding sites in the form of position weight matrices (PWM) and position frequency matrices (PFM) and those containing all the transcription factor binding efficiencies to an exhaustive set of n-mers (n=8-12). The former set are more well-known (JASPAR2008 and Transfac2.0) representations of binding site databases whereas the latter have emerged recently as a product of high throughput methods for screening binding efficiencies. The Uniprobe database (Newburger etal., 2009) is a repository of binding site information of this class generated by universal protein binding microarray (PBM) technology (reference) and is the source for three search databases used by SANN (Berger etal., 2008; Badis etal., 2009; Zhu etal., 2009)

The method of storing binding site information in a position-weight-matrix essentially represents a statistical averaging of several putative and/or functional transcription factor target sites (Prestridge etal., 1993). As such, it cannot retain information about nucleotide covariance at different positions (Benos etal., 2002). While this is an adequate approach to capturing the identity of core nucleotide positions especially when binding site information is sparse, a PWM representation is not a necessary representation when binding site information is abundant or complete (Berger etal., 2008). The major hindrance, however, to scanning all individual binding sites against a query sequence is the high computational time. Through the use of the PASS algorithm search time of a 600bp, 50kb and 200kb input sequences with a dataset of 300,000 10-mer binding sites takes only 2.417, 2.59 and 2.65 seconds respectively. The short runtimes of seed-based search algorithms however, come with a sensitivity tradeoff whereby the risk of missing alignments increases with longer reads (Ma etal. 2002).

SANN generates two types of outputs per inputted sequence. The first is a bird's eye view of the input sequences, annotated to show regions of conservation as defined by reciprocal PASS and BLASTN as well as annotation tracks for each selected database, showing the position and orientation of putative transcription factor binding sites and blast database matches. When more than two sequences are inputted, SANN generates a PASS and BLASTN map for each comparison (resulting in n-1 graphs per sequence) (figure A.2). The second output contains a worm's eye view of the same sequence, zoomed in to the nucleotide level, and containing explicitly the nucleotide sequences of the database matches. Additionally, the sequence is highlighted to show its alignment against other inputted sequences (again, n-1 plots are generated, one for each reciprocal blast alignment) (figure A.3). These primary sequence plots allow the user to identify specific sites for mutation or deletion analysis, as well as view neighboring sequence. The position, orientation and a short description of each sequence annotation are also provided in tabular form.

DESIGN AND IMPLEMENTATION

The SANN web interface is a CGI script coded in Python. The backend codebase is also Python based while the binding site and sequence databases are stored as FASTA flatfiles and blastn databases respectively. The entire codebase is available for download at http://sann.soureforge.net. The codebase makes use of two python packages designed for DNA sequence manipulation (the Seqdb Module of Pygr at http://bioinfo.mbi.ucla.edu/pygr) and motif searching (Motility package at http://cartwheel.idyll.org). Finally, sequence and exhaustive binding site database searches are performed using command-line versions of blast (blastall, http://blast.ncbi.nlm.nih.gov/) and PASS (http://pass.cribi.unipd.it).

CONCLUSION

High throughput sequencing and experimental assays have generated vast resources from which biologists can use to understand living systems, however there is currently a disconnect between the existence of data and the existence of useful tools for probing that data. SANN attempts to bridge that gap between the sequence/binding site databases and the developmental biologist though an accessible, visually intuitive sequence annotation tool. Furthermore, it solves the problem of scanning large databases of explicit binding site sequences by using the PASS algorithm in a novel way. We expect that SANN will be used for organizing and expediting cis-regulatory analysis and for exploring sequence orthologies not only at the level of nucleotide variation but also at the level of binding site cluster membership.

ACKNOWLEDGEMENTS

The author of this work thanks current and former members of the Davidson lab for helpful feedback and discussions.

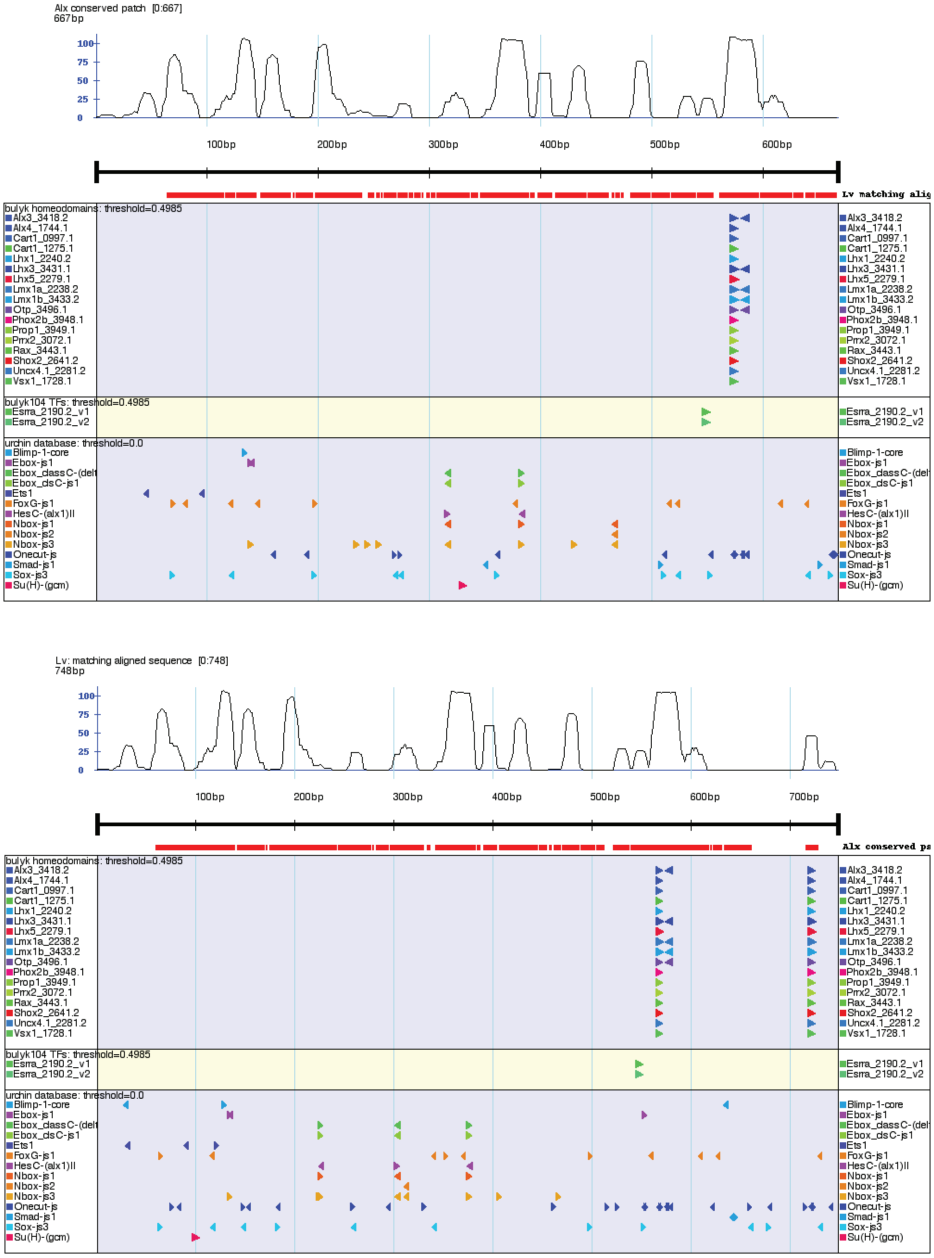


figure1: SANN Bird’s eye view of sequence, showing binding site matches, PASS and BlastN conservation for two orthologous sea urchin genomic sequences (upper: S. purpuratus, lower: L. variegatus). In each graph, reciprocal PASS search is displayed as a histrogram of matches, reciprocal Blastn search as a set of red rectanges, and binding site results as a series of blue/gold tracks.

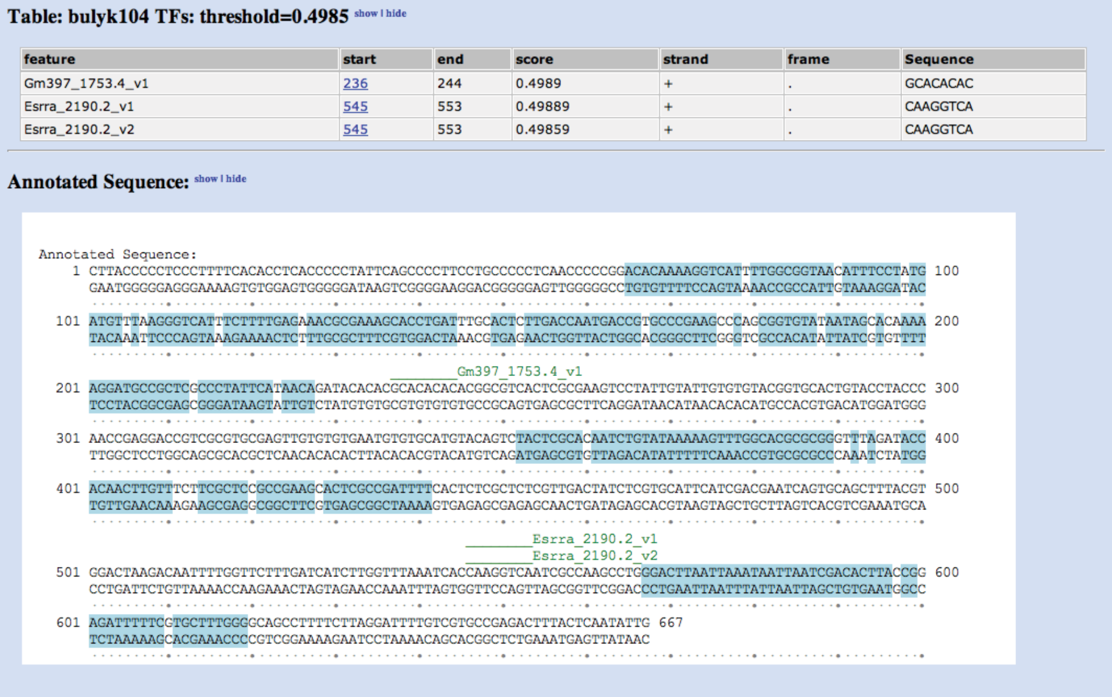


Table: bulyk104 TFs: threshold=0.4985

feature	start	end	score	strand	frame	Sequence
Gm397_1753.4_v1	236	244	0.4989	+	.	GCACACAC
Esrra_2190.2_v1	545	553	0.49889	+	.	CAAGGTCA
Esrra_2190.2_v2	545	553	0.49859	+	.	CAAGGTCA

Annotated Sequence:

```
Annotated Sequence:
  1 CTTACCCCCTCCCTTTTCACACCTCACCCCCTATTCAGCCCCTTCCTGCCCCCTCAACCCCCGGACACAAAAGGTCATTTTGGCGGTAACATTTCCTATG 100
    GAATGGGGGAGGGAAAAGTGTGGAGTGGGGGATAAGTCGGGGAAGGACGGGGGAGTTGGGGGCCTGTGTTTTCCAGTAAAACCGCCATTGTAAAGGATAC

101 ATGTTTAAGGGTCATTTCTTTTGAGAAACGCGAAAGCACCTGATTTGCACTCTTGACCAATGACCGTGCCCGAAGCCCAGCGGTGTATAATAGCACAAAA 200
    TACAAATTCCCAGTAAAGAAAACTCTTTGCGCTTTCGTGGACTAAACGTGAGAACTGGTTACTGGCACGGGCTTCGGGTCGCCACATATTATCGTGTTTT

                                        ________Gm397_1753.4_v1
201 AGGATGCCGCTCGCCCTATTCATAACAGATACACACGCACACACACGGCGTCACTCGCGAAGTCCTATTGTATTGTGTGTACGGTGCACTGTACCTACCC 300
    TCCTACGGCGAGCGGGATAAGTATTGTCTATGTGTGCGTGTGTGTGCCGCAGTGAGCGCTTCAGGATAACATAACACACATGCCACGTGACATGGATGGG

301 AACCGAGGACCGTCGCGTGCGAGTTGTGTGTGAATGTGTGCATGTACAGTCTACTCGCACAATCTGTATAAAAAGTTTGGCACGCGCGGGTTTAGATACC 400
    TTGGCTCCTGGCAGCGCACGCTCAACACACACTTACACACGTACATGTCAGATGAGCGTGTTAGACATATTTTTCAAACCGTGCGCGCCCAAATCTATGG

401 ACAACTTGTTTCTTCGCTCCGCCGAAGCACTCGCCGATTTTCACTCTCGCTCTCGTTGACTATCTCGTGCATTCATCGACGAATCAGTGCAGCTTTACGT 500
    TGTTGAACAAAGAAGCGAGGCGGCTTCGTGAGCGGCTAAAAGTGAGAGCGAGAGCAACTGATAGAGCACGTAAGTAGCTGCTTAGTCACGTCGAAATGCA

                                                 ________Esrra_2190.2_v1
                                                 ________Esrra_2190.2_v2
501 GGACTAAGACAATTTTGGTTCTTTGATCATCTTGGTTTAAATCACCAAGGTCAATCGCCAAGCCTGGGACTTAATTAAATAATTAATCGACACTTACCGG 600
    CCTGATTCTGTTAAAACCAAGAAACTAGTAGAACCAAATTTAGTGGTTCCAGTTAGCGGTTCGGACCCTGAATTAATTTATTAATTAGCTGTGAATGGCC

601 AGATTTTTCGTGCTTTGGGGCAGCCTTTTCTTAGGATTTTGTCGTGCCGAGACTTTACTCAATATTG 667
    TCTAAAAAGCACGAAACCCCGTCGGAAAAGAATCCTAAAACAGCACGGCTCTGAAATGAGTTATAAC
```

Figure 2: A zoomed in view of inputted sequence, showing binding site matches above and below matching nucleotides. Sequence is highlighted to show reciprocal blast alignment. A table of binding site matches, position, orientation and sequence is plotted above the sequence.

REFERENCES

Altschul, S.F., Gish, W., Miller, W., Myers, E.W. and Lipman, D.J. (1990) Basic local alignment search tool. *J Mol Biol*, **215**, 403-410.

Benos, P.V., Bulyk, M.L. and Stormo, G.D. (2002) Additivity in protein-DNA interactions: how good an approximation is it? *Nucleic Acids Res*, **30**, 4442- 4451.

Berger, M.F., Badis, G., Gehrke, A.R., Talukder, S., Philippakis, A.A., Pena- Castillo, L., Alleyne, T.M., Mnaimneh, S., Botvinnik, O.B., Chan, E.T. *et al.* (2008) Variation in homeodomain DNA binding revealed by high-resolution analysis of sequence preferences. *Cell*, **133**, 1266-1276.

Badis, G., Berger, M.F., Philippakis, A.A., Talukder, S., Gehrke, A.R., Jaeger, S.A., Chan, E.T., Metzler, G., Vedenko, A., Chen, X. *et al.* (2009) Diversity and complexity in DNA recognition by transcription factors. *Science*, **324**, 1720- 1723.

Campagna, D., Albiero, A., Bilardi, A., Caniato, E., Forcato, C., Manavski, S., Vitulo, N. and Valle, G. (2009) PASS: a program to align short sequences. *Bioinformatics*, **25**, 967-968.

Lee, P.Y., Nam, J. and Davidson, E.H. (2007) Exclusive developmental functions of gatae cis-regulatory modules in the Strongylocentrorus purpuratus embryo. *Dev Biol*, **307**, 434-445.

Ma, B., Tromp, J. and Li, M. (2002) PatternHunter: faster and more sensitive homology search. *Bioinformatics*, **18**, 440-445.

Newburger, D.E. and Bulyk, M.L. (2009) UniPROBE: an online database of protein binding microarray data on protein-DNA interactions. *Nucleic Acids Res*, **37**, D77-82.

Prestridge, D.S. and Stormo, G. (1993) SIGNAL SCAN 3.0: new database and program features. *Comput Appl Biosci*, **9**, 113-115.

Smith, J. (2008) A protocol describing the principles of cis-regulatory analysis in the sea urchin. *Nat Protoc*, **3**, 710-718.

von Bubnoff, A., Peiffer, D.A., Blitz, I.L., Hayata, T., Ogata, S., Zeng, Q., Trunnell, M. and Cho, K.W. (2005) Phylogenetic footprinting and genome scanning identify vertebrate BMP response elements and new target genes. *Dev Biol*, **281**, 210-226.

Yuh, C.H., Dorman, E.R., Howard, M.L. and Davidson, E.H. (2004) An otx cis-regulatory module: a key node in the sea urchin endomesoderm gene regulatory network. *Dev Biol*, **269**, 536-551.

Zhu, C., Byers, K.J., McCord, R.P., Shi, Z., Berger, M.F., Newburger, D.E., Saulrieta, K., Smith, Z., Shah, M.V., Radhakrishnan, M. *et al.* (2009) High- resolution DNA-binding specificity analysis of yeast transcription factors. *Genome Res*, **19**, 556-566.