Evolution of regulation of ascidian species

Alicia Madgwick

THÈSE POUR OBTENIR LE GRADE DE DOCTEUR DE L'UNIVERSITÉ DE MONTPELLIER

En Biologie - Santé

École doctorale Science Chimiques et Biologiques pour la Santé

Centre de Recherche en Biologie cellulaire de Montpellier UMR 5237

Evolution des programmes transcriptionnels développementaux des ascidies *Ciona robusta* et *Phallusia mammillata*

Présentée par Alicia MADGWICK

Le 7 novembre 2017

Sous la direction de Patrick LEMAIRE

Devant le jury composé de

Sébastien Darras, Dr., UPMC Université Sorbonne	Rapporteur
Benjamin Prud'homme, IBDM UMR 7288	Rapporteur
Mounia Lagha, CNRS-IGMM UMR 5535	Examinatrice
Justin Crocker, EMBL Heidelberg	Examinateur
Patrick Lemaire, CNRS-CRBM UMR 5237	Directeur de thèse

UNIVERSITÉ DE MONTPELLIER

Remerciements

I would like to start off by thanking my supervisor, Patrick Lemaire. Over the course of my four years in the lab, I enjoyed his trust and patience allowing me to have the freedom and the time to discover and pursue my research interests. I feel that I learned a lot through this process.

I would also like to thank my PhD jury members Sébastien Darras, Benjamin Prud'homme, Mounia Lagha and Justin Crocker with a special thank you to Sébastien Darras and Benjamin Prud'homme for accepting to read and correct my thesis and again Benjamin Prud'homme for accepting to be the president of the jury. I would like to thank Sébastien Darras, yet again, along with Emmanuel Faure and Edouard Bertrand for participating in my comité de suivi de thèse over the years.

My time in the CRBM would not have been as much fun without the whole of the Lemaire lab (past and present)! I would especially like to thank Matija for his friendship and companionship when the lab was small, Damien for being so invested in the project and for being great fun to work with and Sabrina, Jacques and Ulla for being so very helpful and friendly. I would also like to thank Marta for teaching me a new technique but also for an enjoyable collaboration.

This Montpellier experience was further enhanced by many people from the CRBM and IGMM. I enjoyed their company at the lab and in the evening with a beer, in particular: Matija, Damien, Fanny, Etienne, Jabran, Emilie, David, Quentin, François, Sabrina, Jörg, Lena, Bruno, Barbara, Madhi, Nicolas. I would also like to thank Thierry for being such a friendly CRBM neighbour.

Finally, I would like to finish by thanking my family. I was very pleased that my parents made the trip from Switzerland to come to my PhD viva. Most of all, I would like to thank Léo whose encouragement and patience gave me a huge support throughout the years. Our scientific discussions throughout my project were invaluable to my progression.

Résumé

Comment la morphogenèse embryonnaire peut-elle être conservée malgré une divergence importante des séquences codantes et non-codantes ? Pour répondre à cette question, nous avons travaillé sur le développement précoce d'ascidies divergentes, *Phallusia mammillata* et *Ciona intestinalis*. Ces espèces partagent une morphogénèse pratiquement identique et des lignages cellulaires stéréotypés. Or, leurs génomes sont tellement divergents que leurs séquences ne peuvent pas être alignées.

Nous avons choisi d'étudier les cellules précurseuses de l'endoderme au cours de deux processus développementaux conservés : spécification du destin et la gastrulation. Nous avons comparé par hybridation *in situ* l'expression transcriptionelle des gènes régulateurs orthologues dans *Phallusia* et *Ciona*. Nous avons trouvé que l'expression dans l'endoderme de 8 gènes régulateurs impliqués dans ces processus développementaux est qualitativement conservée entre les deux espèces.

Pour étudier comment ces gènes ont conservé leur régulation malgré une divergence non-codante importante, nous avons collaboré avec l'équipe Gomez-Skarmeta pour cartographier, par ATAC-seq, la chromatine ouverte dans les deux espèces pour identifier les régions régulatrices actives à l'échelle du génome. 35 sur les 39 séquences ouvertes avoisinant les gènes de l'endoderme ont été trouvé active avant le stade larval, par éléctroporation. La plupart des séquences testées ont conservé leur activité dans les deux espèces malgré la divergence de séquence. Nous avons alors identifié des sites de fixations pour facteurs de transcription potentiels se trouvant dans les enhancers pour l'endoderme pour identifier les régulateurs dans *Phallusia* et *Ciona*.

Nos résultats suggèrent des changements assez importants de l'ordre des sites de fixations sans pour autant avoir de changement dans l'architecture dans les réseaux de gènes régulateurs ; ceci explique la conservation qualitative de l'expression des gènes entre ces ascidies divergentes. En outre, nous avons trouvé que les shadow enhancers sont plus répandus qu'anticipé.

Abstract

How can embryonic morphogenesis be evolutionarily conserved in spite of extensive divergence in coding and non-coding genome sequences? To address this question, we worked on the early development of two very divergent ascidians, Phallusia mammillata and Ciona intestinalis. These species share an almost identical early morphogenesis and stereotyped cell lineages. Remarkably, however, their genomes are divergent to the extent that their non-coding sequences cannot be aligned and gene order has not been conserved.

We focus our attention on the behaviour of endoderm precursors throughout two important evolutionarily conserved developmental processes: initial fate specification and early gastrulation. We first compared by *in situ* hybridisation the transcriptional expression of orthologous regulatory genes in *Phallusia* and in *Ciona*. We found that the endodermal expression of 8 regulatory genes known to be involved in these developmental processes is qualitatively conserved between the two species.

To study how these genes conserved their regulation in spite of extensive non-coding sequence divergence, we collaborated with the Gomez-Skarmeta lab to map, by ATAC-seq, open chromatin regions in both species to identify active regulatory regions genomewide. Three quarters of the 39 open chromatin regions for endodermal genes behaved as active regulatory sequences by the larval stage, when tested by electroporation in embryos. Many of the tested sequences had conserved *cis*-regulatory activity in both species in spite of sequence divergence. We have identifed putative transcription factor binding sites in endodermal enhancers in both species to identify conserved upstream regulators shared between *Phallusia* and *Ciona*.

Taken together our results suggest that extensive transcription factor binding site turn over, without radical change in GRNs architecture, may explain the qualitative conservation of gene expression patterns between highly divergent ascidian genomes. Furthermore, we found that shadow enhancers are much more prevalent than initially anticipated.

Taken together our results suggest that extensive transcription factor binding site turn over, without radical change in GRNs architecture, may explain the qualitative conservation of gene expression patterns between highly divergent ascidian genomes. Furthermore, we found that shadow enhancers are much more prevalent than initially anticipated.

Contents

Introduction

I. Gene regulation

Our bodies are made of cells of many different shapes, sizes, and functions. In spite of this diversity, they all share (with a few rare exceptions) the same set of genes, encoded into the sequence of a very long polymer, DNA (Deoxyribonucleic Acid). How can cells containing an identical DNA content (the genome) differ so much has long been puzzling. It is now understood that what differs inside these cells are the proteins that are synthesised from the genes encoded in DNA. For instance, a red blood cell will be full of the globin protein that carries oxygen, a muscle cells will be enriched in proteins that will allow it to contract, while a neurone will harbour many proteins that transmit electric impulsions.

In his famous 1970 paper "Central Dogma of Molecular Biology", Francis Crick stated that the most common route to producing protein is not directly from the DNA (Crick, 1970). Indeed, DNA is copied into an intermediate messenger molecule: RNA (Figure 1A); this process is called transcription. The first studies into transcription found that different cell types produced different amounts of RNA from their identical copies of DNA (Britten and Davidson, 1969). This process is called gene regulation. Although RNA maturation, its translation into protein, subsequent modifications of these proteins, and the control of their degradation all contribute towards a cell's phenotype, the initial production of RNA from DNA is the first direct read-out of the genome. In this work, we will therefore be focusing on gene regulation.

1. Regulation of transcription

The production of RNA from a gene can be broken down into three main steps all driven by specific sets of proteins: initiation, elongation and termination. Initiation starts with the binding to the DNA of the pre-initiation complex (PIC) composed of general transcription factors that recruit the enzyme that copies DNA into RNA, the RNA polymerase, at the beginning of the gene (Figure 1B). Then, during the elongation step,

the polymerase progresses along the gene, reading the DNA, to produce a single stranded RNA molecule. Once the polymerase reaches a specific signal at the end of the gene, it dislodges from the DNA, releasing a fully transcribed new transcript in the nucleus.

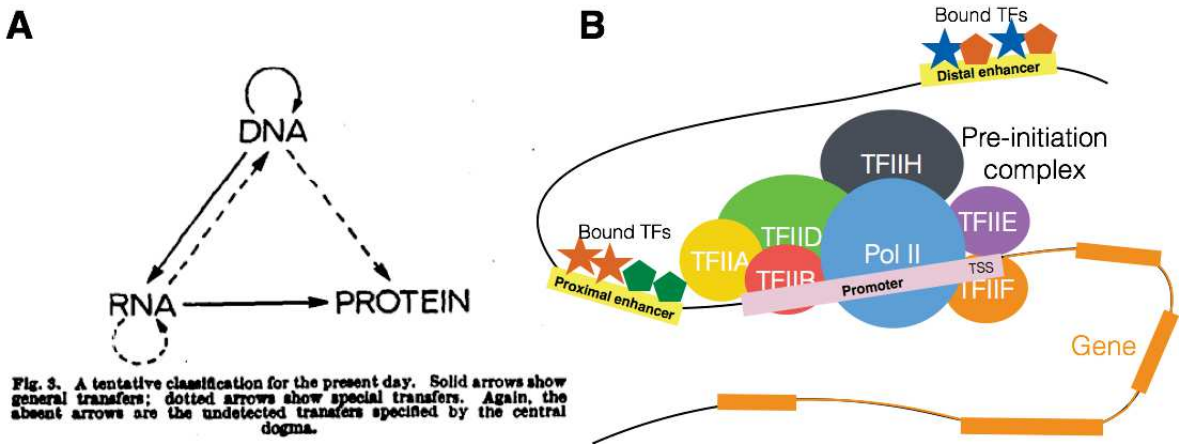


Figure 1. Schematics of transcription. A) Francis Crick's "Central Dogma of Molecular Biology" as of 1970 (Crick, 1970). B) A representation of the initiation of transcription showing the pre-initiation complex (PIC) assembled at the beginning (transcription start site, or TSS) of the gene in orange. The PIC is recruited by other classes of regulatory dNA elements, called "enhancers" found either proximally or distally in yellow, bound by specific proteins, called transcription factors (TFs) represented by stars and hexagons, and interating with the PIC.

In Humans, only 1.22% of the genome encodes protein sequences. The vast majority of the genome is therefore considered "non-coding". For many years, non-coding sequences were not thought to be important and were even referred to as "junk DNA" (Comings, 1972). Between 25% (Graur, 2017) and 80% (ENCODE Project Consortium, 2012) of these non-coding sequences are, however, under the pressure of natural selection or are associated to a biochemical function respectively. They are therefore believed to be necessary for the survival of the species.

Although some sequences controlling transcription are within coding regions (Barthel and Liu, 2008; Lang et al., 2005), the vast majority are thought to be buried within non-coding sequences. In nearly all cases, these sequences are found to be on the same chromosome as the gene they control (scientists in this field say they are "in *cis*"), frequently within a few thousand bases of its transcription start site (TSS), though some *cis*-regulatory sequences can be millions of bases away from the gene (de Kok et al., 1996; Miele and Dekker, 2008; Rebeiz et al., 2009). I will discuss the two main types of *cis*-regulatory sequences: promoters and enhancers. Promoters define the site of recruitment of the PreInitiation Complex, while enhancers are located further away and communicate with the promoter to activate transcription.

2. Promoters

The promoter of a gene is located around the site of transcription initiation. It is estimated that there are about 70,000 promoters in the Human genome (Dunham et al.,

2012), suggesting that many genes have more than one promoter. The first promoters where identified in the mid-1970s in bacteria, which are devoid of nuclei, and just a few years later in eukaryotes, in which genomic DNA is generally larger and always encased in a nucleus (Dhar et al., 1974; Gannon et al., 1979; Pribnow, 1975). The first eukaryotic regulatory sequence to ever be identified contained a so-called TATA box, known at the time as the Goldberg-Hogness box (M.L. Goldberg, PhD thesis, Stanford, 1979). Only around 24% of characterized Human promoters, however, contain a TATA box (Yang et al., 2007). This motif is preferentially found in promoters controlling tissue specific expression, and is associated to the presence of a unique strong TSS (Haberle and Lenhard, 2016). Some of the remaining core promoters harbor a combination of other motifs, including Initiator (Inr), Downstream Promoter Element (DPE), TFIIB recognition element (BRE), Downstream Core Element (DCE), Motif Ten Element (MTE) or TCT motif, but many include none of these elements (Figure 2) (Haberle and Lenhard, 2016).

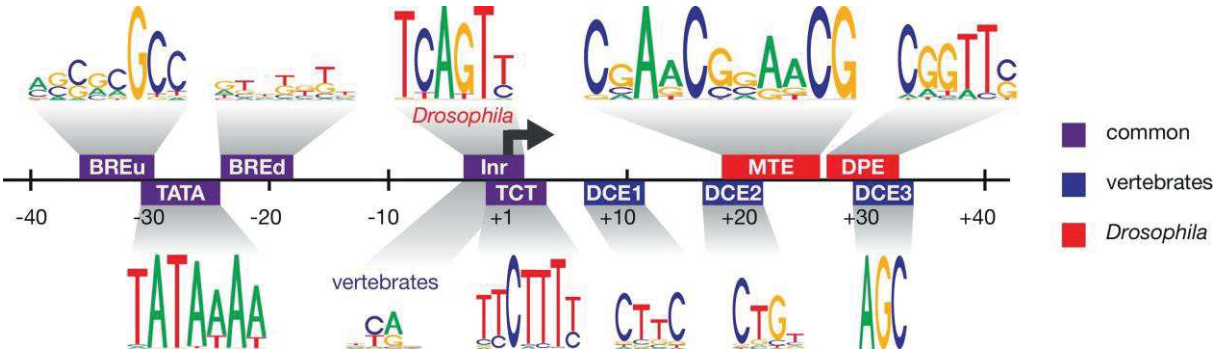


Figure 2. A schematic demonstrating the different promoter motifs (Haberle and Lenhard, 2016).

Some features of promoters also appear to be species specific. For instance, around 70% of vertebrate promoters are located within CG-rich regions, called CpG islands (Deaton and Bird, 2011), a class of genomic region which does not exist in many invertebrates. Computationally identifying promoters on a global scale solely based on their sequence remains highly inefficient. Fortunately, other signatures exist that facilitate the experimental identification of promoters at the genome scale. First, promoters overlap with TSS, which can be experimentally mapped with 1bp resolution by sequencing RNAs. Two classes of TSS were identified. Some promoters drive transcription from "sharp" TSS precisely-positioned to the single base pair, while others drive transcription from “broad” or “dispersed” promoters, equally using multiple TSSs distributed across up to 100 bases (Carninci et al., 2006). Sharp promoters are preferentially associated with TATA boxes and drive the expression of tissue specific or regulatory genes. Broad promoters are associated to CpG islands in vertebrates and mostly control housekeeping ubiquitously expressed genes (Haberle and Lenhard, 2016).

Second, DNA in the nucleus is not naked. It is wrapped around specific proteins called histones, forming what is called chromatin. The histones can be modified after their synthesis, and these modifications affect the compaction of chromatin and the

accessibility of DNA to transcription factors or the PIC. Trimethylation (the addition of three methyl groups) of lysine 4 in histone H3 (H3K4me3) (Heintzman et al., 2007) is found in nearly 75% of human promoters, independently of the expression level of the corresponding gene (Guenther et al., 2007). This latter property came as a surprise and contributed to the realisation that expression of most genes is controlled at the level of transcription elongation rather than transcription initiation, as previously thought, and that RNA Polymerase II is frequently recruited but paused on the promoters of many genes prior to their activation (Adelman and Lis, 2012; Kwak and Lis, 2013). Not all promoters are, however, marked by H3K4me3, the promoters of regulatory genes often lacking this canonical pattern (Pérez-Lluch et al., 2015).

3. Enhancers and Transcription factors

Francis Crick once commented that following a new scientific discovery, it becomes difficult, if not impossible, to imagine in ones mind how things had been previously perceived. This is particularly true with regards to the discovery of enhancers by Sandro Rusconi from Schaffners lab in 1980. This discovery revolutionised our interpretation of development. It is no longer possible to imagine the complexities of development without knowing about *cis*-regulatory sequences and their responsibility in orchestrating gene expression so precisely throughout. Funnily enough, Rusconis has not always been seen as such a revolutionary discovery. When the existence of enhancers was first introduced at a conference in Berlin in the summer of 1980, few people seemed interested in the "enhancing" effects of the DNA sequence in SV40, to Schaffners great disappointment (Schaffner, 2015).

It has been estimated that there are hundreds of thousands of enhancers in the mammalian genome, therefore, roughly 10 enhancers for every gene (Birney, 2012; Shen et al., 2012). Why are there so many enhancers and what is their importance in development?

i. Defining enhancers and transcription factors

Enhancer sequences have been defined as *cis*-regulatory elements that enhance the transcription of a gene (Blackwood and Kadonaga, 1998). Enhancers function as binding platforms for transcription factors (TFs), and contain specific transcription factor binding sites (TFBSs) (Aza-Blanc et al., 1997). TFs are the dynamic component of the enhancer activity. They bind enhancers (and sometimes promoters) in a sequence-specific manner and interact with a variety of cofactors including general transcription machinery proteins. They can either act as activators of transcription, or as repressors, depending on the specific TF but also on the cellular context (Stampfel et al., 2015).

Despite the large number of different TFs - ranging from around 300 in E/coli to over 2000 in humans - the large majority of eukaryotic TFs belong to only around 30 structural families based on the amino acid sequence and 3D architecture of their DNA binding domain (Wingender et al., 2013). TFs have a modular architecture and contain one or more DNA-binding domains and one or more protein interaction domains; the first allows them to bind to TFBSs (around 5-12bps) and the latter to other transcriptional regulators.

The presence on the enhancer of TFs that will either activate or inhibit its activity defines the expression pattern of the gene. Combinations of different activators and repressors can create very specific expression patterns as observed, for instance, in the 2nd band of embryonic expression of the *Drosophila evenskipped* gene, which is controlled by the transcriptional activators Hunchback and Bicoid but also by the short-range transcriptional repressors Giant and Krüppel (Gray and Levine, 1996).

Finally, the definition of an enhancer as stated above was perhaps a bit too restrictive. Enhancers are not the only regulatory sequences to recruit TFs; the proximal promoter just 200bp from the TSS can also act as an enhancer by binding TFs (Ohler and Wassarman, 2010). Additionally, an enhancer can act as a promoter by recruiting RNA Polymerase II to transcribe itself. These enhancer transcripts are known as eRNA, which are thought to help stabilise TF binding and gene transcription (Lai and Shiekhattar, 2014; Lam et al., 2014). This class of enhancers can be identified simply by short RNA-seq.

ii. Enhancer location

On average, a promoter is regulated by 5 enhancers and an enhancer controls two promoters in mammals (Arensbergen et al., 2014). The distance between an enhancers and the promoter(s) of the gene it activates is very variable, ranging in the large vertebrate genomes from a few tens of nucleotides to millions of bases (Lettice et al., 2002). The distance of the enhancer from its target gene can vary quite drastically. Often enhancers are located just upstream or downstream of the gene or even within the intronic regions; however, there are some cases where enhancers can be found much further away. Such examples include an enhancer regulating *cut* in drosophila which was found 80kb upstream of its promoter (Jack et al., 1991) or even an enhancer regulating *Shh* in mouse found 1Mb away within the intron of another gene no less (Lettice et al., 2003). On average, enhancers are located around 100kb from their target promoters in mammals (Jin et al., 2013). In rare cases, enhancers have even been found to be located on a different chromosome to their target gene, in which case their activity is defined as *trans*-regulatory (Gohl et al., 2008).

Enhancers are believed to act by interacting with their cognate promoter and several models have been put forward to permit this interaction, including looping (Krivega and Dean, 2012; Zhang et al., 2013). Some enhancers act hundreds of kilobases away from their target genes, and do not activate genes that are often located between them and their target (Zhang et al., 2013). The enhancer-promoter interaction network, which varies with cell types, is substantially more complex than estimated before chromatin conformation capture techniques were developed (Arensbergen et al., 2014). The main principles defining the range of action of a given enhancer and its specificity for its target promoters remain, however, poorly understood; though we know that it may involve specific interactions between enhancer-bound TFs and promoter-bound proteins, insulator elements and proteins such as cohesin which contribute to large-scale nuclear chromatin architecture and the definition of so-called topologically-associated domains (TADs) within which enhancer-promoter are facilitated (Arensbergen et al., 2014). Finally, it should be pointed out that the distinction between core promoters and enhancers may not be as clear cut as we initially thought: some

promoters bind transcription factors and some can even act as long range enhancers (Catarino et al., 2017).

4. Techniques for mapping regulatory activity

There has been an explosion of techniques to identify regulatory elements. The tools range from understanding the 3D conformations of chromatin to the binding of TFs on enhancers. These tools can help elucidate the regulatory behaviours that unfold within a cell or an organism (Figure 3).

i. Genomewide identification of enhancers

Enhancer sequences can be characterised by their local chromatin structure. A local enrichment of certain histone modifications such as mono-methylation of lysine 4 on histone 3 (H3K4me1) and acetylation of lysine 27 again on histone 3 (H3K27ac) (Bulger and Groudine, 2011) is a frequent hallmark of enhancer activity. ChIP-seq for specific histone modifications, including H3K27Ac and H3K4me1, has allowed the large-scale identification of hundreds of thousands of mammalian promoters and enhancers. Additionally, enhancers are often found in open chromatin regions; this will be discussed later on (Shlyueva et al., 2014).

Techniques to map open chromatin regions at the genome scale have proven invaluable to identifying regulatory regions. DNA compaction and opening is due to nucleosomes. Nucleosomes are 147bps of DNA sequence wrapped around a histone octamer that creates the building blocks for compacted DNA known as heterochromatin. In this situation, genes are inactive. When the nucleosomes are more dispersed due to nucleosome remodelling enzymes and the binding of proteins to the DNA (TF, PIC and RNA Pol II) (Struhl and Segal, 2013), the chromatin becomes open, known as euchromatin, which is associated with active sequences and genes.

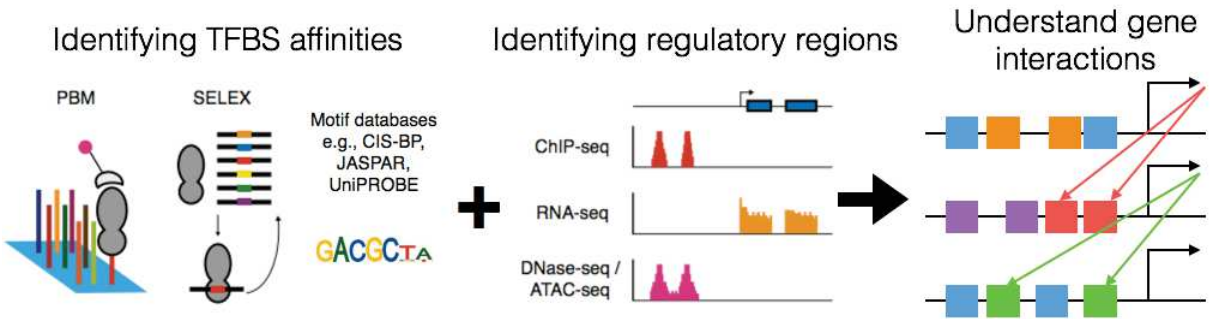


Figure 3. Combining several tools to elucidate regulatory logic: identifying TFBS, by PBM or SELEX, within enhancers than have been mapped, by ChIP-seq and ATAC-seq, could help understand gene interactions (adapted from Inukai et al. 2017).

The combination of tools identifying TFBSs such as SELEX and identifying regulatory regions such as ATAC-seq can be used to build the regulatory landscapes of the genes of interest (Figure 3) and through this help decipher the networks of regulation between each of these genes. Functional chromatin regions have been mapped genomewide such as in the ENCODE projects (Encyclopedia of DNA Elements) in humans, mouseENCODE in mouse and modENCODE in drosophila and worm (ENCODE Project Consortium, 2012; Shen et al., 2012). Several approaches focus on localising regulatory sequences by

identifying nucleosomes or open chromatin by ATAC-seq, DNase-seq and FAIRE-seq and histone modifications by ChIP-seq.

ii. 3D regulation

The chromatin 3D architecture is also a popular topic of interest with regards to gene expression as the physical properties bringing together the enhancer and the promoter, for example, are to be considered. The looping model is vastly promoted, demonstrating how an enhancer can act from a distance (Kleinjan and van Heyningen, 2005).

Techniques such 5C or Hi-C map physical interactions by capturing the 3D chromatin conformation (or the landscape of enhancer-promoter interactions) genome-wide. These sorts of techniques have revealed the existence of multi-layers of 3D conformation involved in gene regulation (Phillips-Cremins, 2014). The looping interactions are organised into megabase-scale TADs (Topology Associated Domains) covering 90% of the genome which then organise themselves into specific sub-territories (Dixon et al., 2012). These structures are thought to bring genes that are co-expressed closer together to create what is known as transcription factories in precise areas of the nucleus (Gibcus and Dekker, 2013).

iii. Enhancer activity

There are many techniques that allow the identification of active enhancers. A classical example is by using a reporter construct by cloning the enhancer in front of a reporter gene (GFP or LacZ encoding β-galactosidase), which is then inserted *in vivo* into the organism. The spatiotemporal expression of the reporter gene can be visualised by microscopy. It is also possible to quantify the enhancer activity by enhancer-FACS-seq (eFS or enhancer Fluorescent Activated Cell Sorting-sequencing) which sorts the cells by their fluorescence (Gisselbrecht et al., 2013); alternatively, a barcode can be added to the transcribed region followed by mRNA sequencing to quantify the level of enhancer activity (Farley et al., 2015; Kheradpour et al., 2013; Nam and Davidson, 2012).

These quantitative techniques cannot give a spatial profile of enhancer activity. The idea behind these techniques is very similar to STARR-seq (Self-Transcribing Active Regulatory Region Sequencing), which is to map active enhancers genome-wide by hooking large number of candidate enhancers to diagnostic or bar-coded transcribed sequences, which upon RNA-seq reveal which candidate enhancer is active. Surprisingly, Arnold *et al.* found that a third of the identified enhancers are normally not active in those cells (Arnold et al., 2013), possibly because STARR-seq subtracts them from the influence of a silencing local chromatin environment at their endogenous locus.

iv. Identifying TFBSs

Several *in vivo* and *in vitro* methods have been developed to elucidate the binding sequences of each TF (see review Levo and Segal, 2014). *In vitro* techniques often work by revealing protein binding to short DNA sequences such as EMSA, PBM and SELEX. Electrophoretic Mobility Shift Assay (EMSA) views this interaction one TF at a time by gel migration to analyse the number of bound TFs. SELEX (Systematic Evolution of Ligands by EXponetial enrichments), however, can be used when nothing is known about the TF sequence affinity. Bound sequences are selected by pull down and then amplified by PCR followed by sequencing to collect binding affinities of a large number

of TFs at once. *In vivo* techniques use immunoprecipitation of chromatin fragments such as ChIP-on-chip or more recently ChIP-seq, which immunoprecipitates the DNA fragment bound in vivo by a given TF. *In vivo* and *in vitro* binding specificities have in general been found to give similar results (Orenstein and Shamir, 2014).

There are several ways to quantitatively represent the DNA-binding specificity of a TF, including position weight matrices or a logo sequence. DNA-binding specificity data generated across different organisms have been regrouped in databases such as Jaspar (Mathelier et al., 2016). This work highlights that binding site specificity is greatly conserved during evolution so that data from any metazoan species can frequently be used to find sites within distant metazoa (Nitta et al., 2015). However, there are some exceptions as TF binding specificity has been known to be subject to adaptive change (Lynch and Wagner, 2008; McKeown et al., 2014; Sayou et al., 2014). This extensive atlas is a great tool to identify TFBSs within known enhancers; this has been done at a larger scale in drosophila using thermodynamic models (Segal et al., 2008). Inversely, using these tools to look for clusters of TFBSs has also successfully been used to find enhancers (Khoueiry et al., 2010; Markstein et al., 2002; Roure et al., 2014).

5. The significance of regulatory sequences

The advancement of technology is helping to improve our understanding of regulation. Here, I will focus on discussing what is known about enhancers and TFBSs.

i. Enhancer syntax

Each enhancer has several binding sites for different TFs and the combination of these TFBSs will determine the activity of the enhancer (Arnone and Davidson, 1997). There are three proposed models for enhancer information processing: the "enhanceosome", the "flexible billboard" and the "TF collective"; all vary in TFBS organisation and flexibility (Arnosti and Kulkarni, 2005; Junion et al., 2012). In the enhanceosome, a strict layout and organisation of the TFBSs is imperative for the correct cooperation of all of the bound TFs to act in unison as a single regulatory unit (Panne, 2008; Panne et al., 2007). The constraint to maintain the same order of TFBSs could explain why regulatory sequences are often more conserved than other non-coding regions across divergent species such as human and mouse (Harmston et al., 2013). However, this enhancer model does not coincide with what is most commonly observed. In many homologous enhancers found in distantly related species, TFBSs seem to change positions within the same region, known as TFBS turnover. This has been demonstrated in the enhancers of *even skipped* across the *Drosophila* family (Figure 4) (Hare et al., 2008). The "flexible billboard" model proposes that enhancers do not need to conserve TFBS order to maintain the same activity.

In truth, these first two models could represent the two ends of the enhancer logic spectrum. Eileen Furlong's lab proposed the TF collective model that depends on protein-DNA and protein-protein binding to coordinate TF binding to the enhancer. This model also allows for a flexible order of TFBSs (Junion et al., 2012).

Papatsenko and Levine argue that the extent to which an enhancer sequence is constrained is related to the TF concentrations; if the concentrations are low, the

sequence is more constrained and vice versa (Papatsenko and Levine, 2007). Not all enhancers, however, have conserved their activity. In fact, thousands of mammalian enhancers are highly evolvable due to the positive selection of certain genes (Villar et al., 2015).

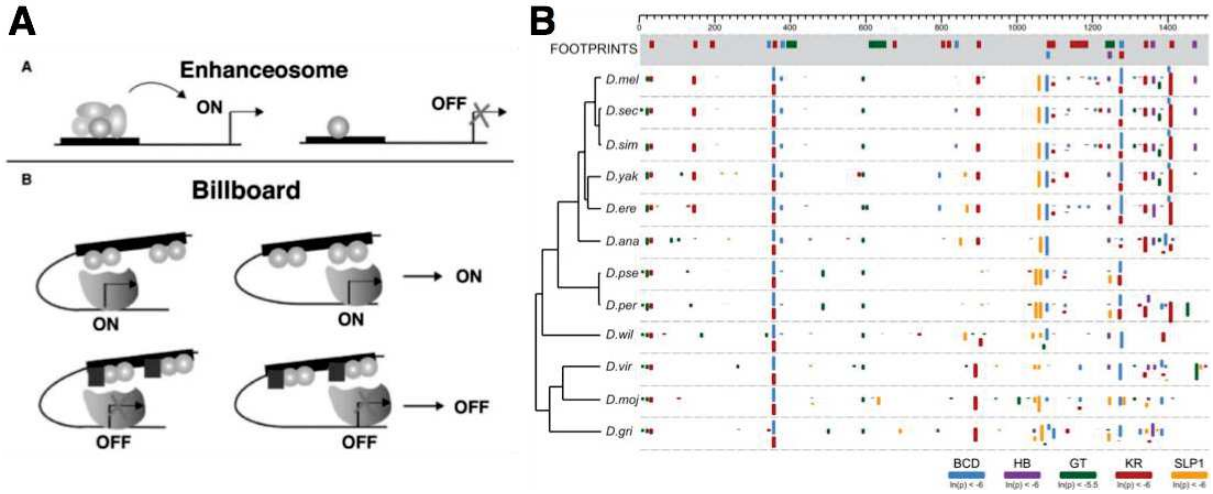


Figure 4. Enhancer logic. A) A representation of the "enhanceosome" vs the "flexible billboard" (Arnosti and Kulkarni, 2005). B) TFBS turnover in the enhancers driving even skipped across Drosophila.

ii. TFs and their binding sites

Enhancers are made up of a cluster of several TFBSs for several different TFs (He et al., 2011). The many experimental tools mentioned above have revealed the implications of these clusters of TFBSs within evolution. For instance, in some cases, TFBSs and more specifically their binding affinities were found to be very important to control expression. Binding sites within *Drosophila* and ascidians were found to be sub-optimum to increase specificity of the expression territories (Crocker et al., 2015; Farley et al., 2015). Through time, the shuffling of position of TFBSs, known as turnover, has also been found whilst maintaining the same enhancer activity (Khoueiry et al., 2017; Rebeiz et al., ///).

The importance of each of the TFBSs within an enhancer is linked to the activity of the TFs. The first TFs to bind to the enhancer are in charge of opening the chromatin by removing the nucleosomes; these are known as pioneering factors (Zaret and Carroll, 2011). Additionally, certain TFs prefer to act together, as partners, to bind to neighbouring sites perhaps to better respond in a specific manner to a signal that is itself less specific (Inukai et al., 2017; Mullen et al., 2011; Trompouki et al., 2011). However, not all TFBSs are necessarily needed for normal development and can be functionally redundant. In the case of the 800bp enhancer driving expression of the 2nd even-skipped band, 300bp can be deleted without consequences in normal conditions even though this sequence contains functional TFBSs. These additional BSs act as a buffer to maintain robustness of gene expression to varying environmental conditions (Ludwig et al., 2011).

iii. Exception to the rule

A gene that is repeatedly expressed throughout development is thought to have several different enhancers controlling it at different times; each enhancer is seen as a module that is limited to regulate expression in a specific tissue at a specific time (Arnone and Davidson, 1997), much like the example of *even skipped* seen earlier. However, this does not suffice to explain the high number of regulatory sequences. Analyses of regulatory landscapes have repeatedly revealed the presence of enhancers with overlapping or redundant activity. These functional redundancies have been suggested to increase robustness to genetic or environmental variations (see review Lagha et al., 2012). This has been demonstrated by the deletion of the redundant enhancer, also known as a shadow enhancer driving *shavenbaby,* which at optimal temperature has no detrimental effect, but at sub-optimal temperatures causes developmental defects in the tricomes (Frankel et al., 2010). Many of these redundant enhancers were initially found around developmental genes using "enhancer traps", a tool that inserts a promoter and a reporter gene into the genome to detect enhancers and their activity (Kikuta et al., 2007), and later by Chromatin immunoprecipitation (Perry et al., 2009).

Dating back as far as King and Wilson's work showing that humans and chimpanzees proteins are practically identical, it has been known that gene expression plays a major role in evolution and morphology (King and Wilson, 1975). Many researchers, such as Carroll, have argued the contribution of regulatory evolution to morphological diversity (Carroll, 2008). The regulatory landscapes and the characteristic modularity of enhancers have allowed these cis-regulatory sequences to be responsible for morphology evolution; this has been identified for instance by QTL (quantitative trait loci) (Maurano et al., 2012). However, gene expression is not always finely tuned for a specific developmental process. This can be seen during mouse limb development, *Lunapark* is controlled by the same regulatory landscape as *HoxD* (Spitz et al., 2003) despite the fact that it is not essential for this process.

II. GRNs drive development

So far, transcription and regulation of gene expression have been discussed at the cellular level. While it is now known that gene expression drives development and morphogenesis, how can an activity regulated at the level of individual cells coordinate the development of an entire multicellular organism?

Animal development is driven by a defined genetic toolkit responsible for the specific spatial and temporal formation of each feature of the animal body. A rather small fraction of genes in any given animal make up this toolkit responsible for the patterning and formation of the body through gene expression. These genes are part of one of two groups: transcription factors that control the expression of genes in individual developing cells, and signalling pathways components that control interactions between cells.

1. Evolutionary developmental biology

Early embryologists found that regions of the early embryo were predestined to form a certain tissue or body part. This led to the realisation that there must be factors responsible for their formation well in advance.

i. The Spemann Organiser: cell-cell signalling organises the surrounding tissues

Even before the first genes orchestrating development had been characterised, early embryologists tried to comprehend development starting with how the organisation of the early body plan into the different germ layers (endoderm, ectoderm and mesoderm) and body parts is put in place. This framework proved useful as a starting point for how the organisation of body parts changes throughout development.

In 1924, Hilde Mangold, working with Hans Spemann, transplanted a tiny region located on the dorsal side of the early embryo of a pigmented salamander newt, the lip of the blastopore, to the ventral side of an albino newt embryo of the same age. The transplantation led to the formation of a Siamese-twin structure (Figure 5A). This structure was mostly constituted of unpigmented host cells, revealing that the graft had instructed them to change their original ventral fate to a more dorsal one. This was the first demonstration of an organiser: a tissue or region that has an effect on the fate of the surrounding tissues. Structures homologous to the Spemann organiser have since been found in many other organisms: other amphibians, including *Xenopus* (Figure 1B), fish, birds, mammals (Beddington, 1994) and even the invertebrate chordate *Branchiostoma lanceolatum (Le Petillon et al., 2017)*. Transplantation between *Xenopus* and chick revealed that the organiser signal(s) were intelligible between species (Kintner and Dodd, 1991). The organiser is now considered to be a feature originating at the onset of chordate evolution and now shared by most chordates although not ascidians.

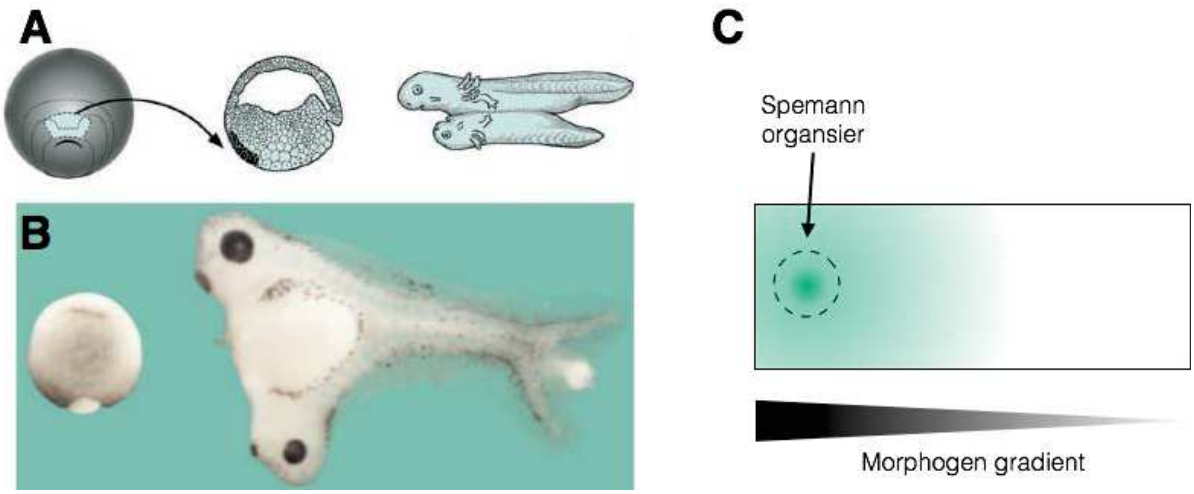


Figure 5. Mangold and Spemann's transplant experiment brought to light the Spemann organiser. A) A schematic of the transplantation in the salamander newt (Carroll et al., 2001). B) The Spemann organiser seen in Xenopus (De Robertis, 2006). C) Schematic showing the morphogen gradient decreasing further from the Spemann organiser.

One explanation for the long-range inductive effect of the Spemann organiser was that it is a source of inducing molecules, or morphogens, whose concentration within a cell or tissue will vary and to which neighbouring cells will respond. These morphogens are interpreted based on the gradient of their concentration that decreases gradually in cells further from the organiser cells (Figure 5C). These different concentrations will induce the different fates of the surrounding cells.

Following the discovery of organisers and morphogens in the first half of the 20th century, further understanding came many years later with the discovery of the genes whose products were found to act as morphogens. These genes were finally recognised as the genetic toolkit of development (Carroll et al., 2001).

ii. The toolkit driving morphogenesis

The first hurdle to studying embryonic development and morphogenesis was identifying which genes drive these processes and how these genes are coordinated from the single egg up to the adult stage composed of several different tissues.

Of the thousands of genes that make up animal genomes, most are either housekeeping genes (involved in essential cellular functions shared by all or most cells of the body) or genes encoding proteins, which confer their specialised, one also says differentiated, functions to particular adult or larval cells or tissues. However, the genes that make up the developmental toolkit are a different set of genes whose role is to determine the body plan and the development of each separate feature or body part by coordinating patterning events. These are mainly TFs and genes involved in signalling pathways. The most fruitful early approach to isolating genes and pathways involved in normal development was by naturally occurring random genetic mutations affecting adult morphology. These viable abnormal organisms were kept for further analysis creating a collection of organisms with different phenotypes. And so began the pioneering work in insects where systematic screens discovered genes involved in development.

This was then achieved by creating random mutations in toolkit genes was through mutagens (chemical or radiation). The most notable work was done in *Drosophila melanogaster* and earned Christiane Nüsslein-Volhard and Eric Wieschaus the Nobel prize in 1995 for their work published in 1980 (Figure 6A). These genetic catastrophes were kept for intensive screens to decipher which genes were responsible for many developmental features (Wieschaus and NV 2016), such as ey gene part of the Pax6 gene family (Figure 6B). This work and many of its kind brought to light homeobox (Hox) genes crucial for body plan shared across many species (Figure 6C).

The subsequent cloning of these genes revealed three major surprises: 1) these genes encode transcription factors and signalling genes. 2) They defined new families of genes, such as the homeobox family, rather than unique genes. 3) Most surprisingly, these genes were not unique to insects but were also found in vertebrates. In fact, many of these genes, such as the Hox genes, have conserved their functions in flies and mammals (Duboule and Dollé, 1989; Graham et al., 1989). The discovery that Hox genes were conserved, between mammals and drosophila no less, was an incredible moment in the

history of developmental biology. There is a much larger signature of unity of developmental processes among metazoa than anticipated.

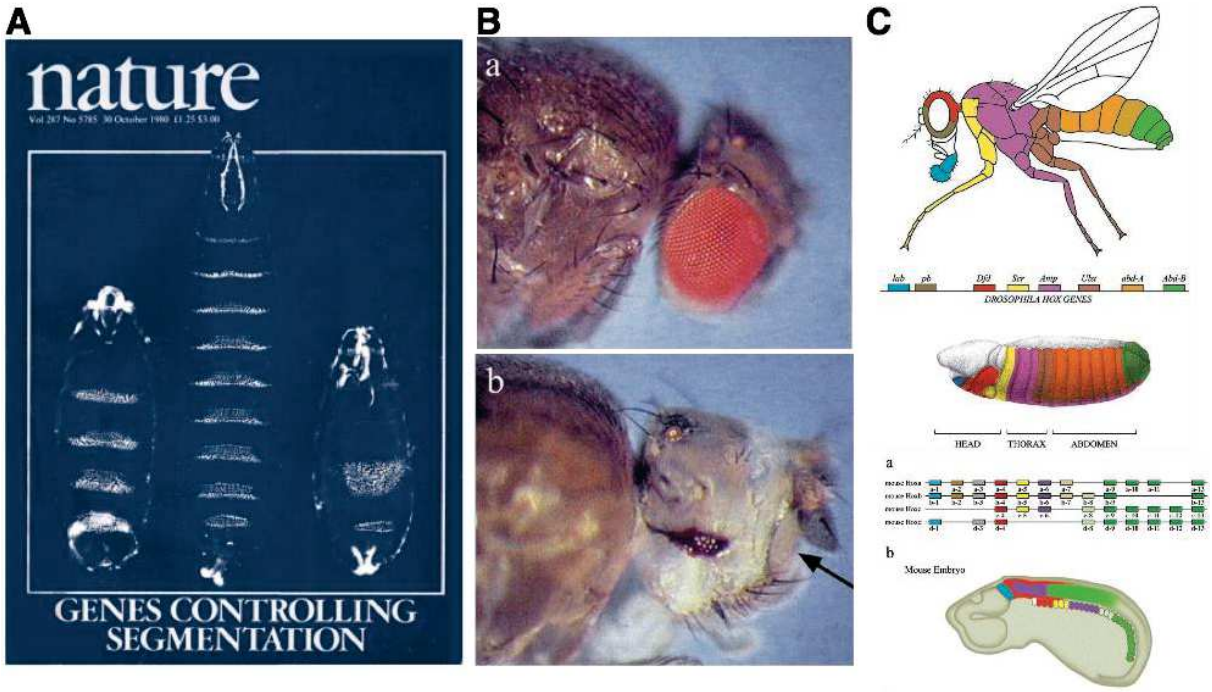


Figure 6. The discovery of developmental genes through the analysis of induced or naturally-occurring Drosophila mutants. A) Eric Wieschaus and Christiane Nüsslein-Volhard's work made the front cover of Nature in 1980. B) Wild type Drosophila eye (a) compared to the ey Drosophila eyeless mutant. C) A schematic of the Hox genes responsible for the different body parts in Drosophila and mouse (Carroll et al., 2001).

Comparing species and the conservation of developmental genes making the same body plans was a first step towards understanding morphogenesis. This was mainly based on descriptive work. Genome sequencing at the turn of the millennium brought comparative genomics to understand the evolution of the genetic toolkit in development. Studying the individual genomes revealed characteristics such as gene number, gene order, gene duplications, repeated sequences and *cis*-regulatory sequences. Gene duplication and the accumulation of sequence change revealed that this could lead to a functionally divergent gene. This opens the door to many interesting questions: was genome duplication, and therefore the expansion of the toolkit, necessary for the emergence and evolution of morphological complexities? Or are there other possible processes that can lead to morphological diversity?

2. Introduction to gene regulatory networks: the organisers of development

The expansion of the toolkit may indeed be a source of morphological diversity. However the answer is not so simple. It was previously mentioned that many of the toolkit genes are shared across species; the difference, therefore, lies in how this toolkit is actually used and deployed. We must first focus on understanding how the toolkit is

used in the development in one species to then understand how it can be tweaked in another.

i. Defining gene regulatory networks

A rather incredible biological phenomenon is development: complex organisms composed of multiple tissues and body parts arise from a single cell, the fertilized egg. This cell must contain not only all of the genetic information necessary for the process to unfold, but also the initial conditions that will initiate the developmental program. The genetic information is contained within the DNA, the initial conditions are maternally deposited factors and the developmental program is driven by the networks of the developmental toolkit genes. How can the same genetic information be interpreted differentially across cells and tissues to give different gene expression (Ben-Tabou de-Leon and Davidson, 2007)?

This expression is orchestrated by the regulatory apparatus, which is made up of two parts. The first is the regulatory genes; TFs that bind DNA to activate or repress gene expression and signalling pathways that carry out inter- and intracellular communications. The set of active TFs within the nucleus is defined as its regulatory state. The second is the *cis*-regulatory regions; these are part of the regulatory genome, which includes regulatory genes. The regulatory genome is the same for each cell of the organism (Figure 7A).

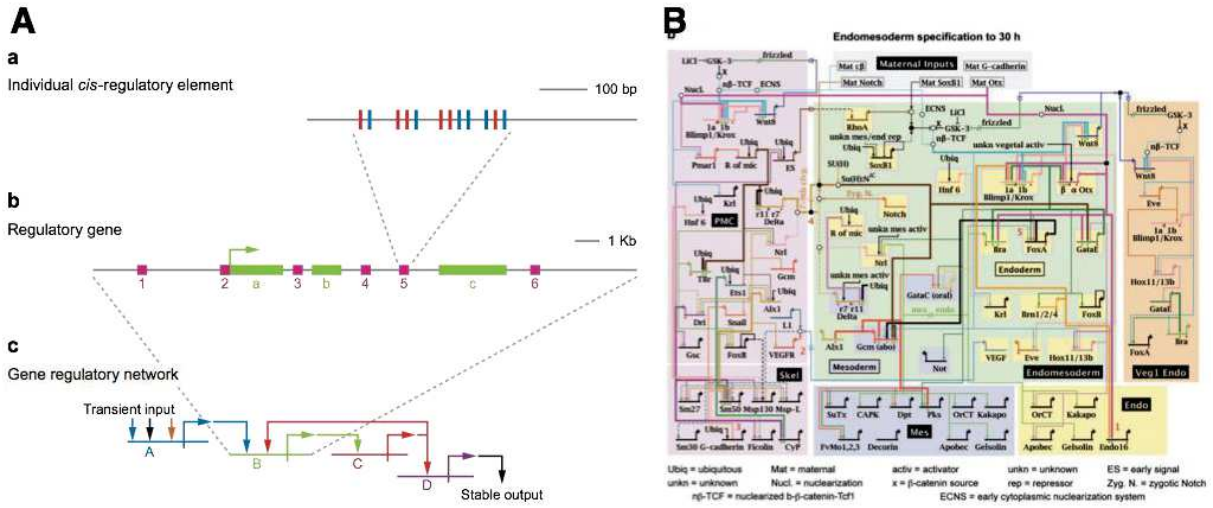


Figure 7. The different components of the transcriptional regulatory genome (Aa) An individual cis-regulatory element with a cluster of two types of binding sites in red and blue. (Ab) A regulatory gene in green with 3 exons (a to c) surrounded by 6 regulatory elements in pink (1 to 6). (Ac) A map of a simple gene regulatory network where by gene A in blue activates B in green by activating the cis-regulatory element of B, leading to B activating C and so on. (B) Schematic representation of the sea urchin GRNs driving endomesoderm development up until 30hpf (Ben-Tabou de-Leon and Davidson, 2007).

The regulatory state of the cells will also define the regulatory state that their progeny will assume. The cells will go through a series of regulatory states; the previous regulatory state will activate the *cis*-regulatory regions of the next set of genes. This is

the process of specification; the cells acquire an identity that is transferred to their progeny. This passing-on of information through the inter-regulating regulatory genes and their *cis*-regulatory regions forms the regulatory networks. These gene regulatory networks, GRNs for short, are the underlying drivers of the developmental program; and therefore to understand development or the evolution of development, we need to build a map of the GRNs (Ben-Tabou de-Leon and Davidson, 2007).

ii. Deciphering the developmental GRNs

Elucidating the GRNs demands a combination of extensive small- and large-scale experimental data. The first step is to identify candidate regulatory genes, which can be done by genome sequencing and annotation. Next, the regulatory states of the cells or tissue need to be defined by temporal and spatial expression data. Because of the cellular heterogeneity of developing embryos, this was until recently mostly performed by whole mount in situ hybridisation (ISH), a technique that could gradually be superseded by single-cell RNA sequencing approaches (Yuan et al., 2017). Finally, the connection between genes needs to be elucidated. This has very often been done by labour-intensive TF perturbation experiments, such as the work done in *Ciona* (Imai et al., 2004, 2006). The most efficient way to definitely confirm the direct regulatory link between two genes is by *cis*-regulatory analysis and mutation.

One of the most well known developmental GRNs drives endomesoderm differentiation and specification in the sea urchin embryo up to the onset of gastrulation about 30 hours post-fertilization under standard laboratory conditions (Figure 7B). From works like this, it is possible to extract general rules of GRNs. The structure and function of different GRNs seem very diverse, however, they all share a set of general features: 1) the activation or repression of each *cis*-regulatory region necessitates a specific combination of TFs; one TF is not sufficient. Therefore, the same TFs can be used in various combinations to define several different specification events. 2) The networks can be broken down into “subcircuits”, which each have their own developmental task. 3) If the regulatory states need to persist for specification, these are maintained by positive feedback loops. If a boundary needs to be defined, this is done by repressing an alternative fate. If a subcircuit needs only to be transiently active, this can be controlled by a negative-feedback loop (Ben-Tabou de-Leon and Davidson, 2007; Peter and Davidson, 2011a, 2011b).

iii. Developmental evolution and GRNs

Once a well-defined map of the GRNs is built, it is possible to understand the molecular mechanisms driving development. To have a detailed map is already challenging, however, to study the evolution of developmental GRNs, it is necessary to have the detailed GRNs of at least two species. To distinguish the direction of evolutionary change, then the GRNs of three or more species are needed. Furthermore, to efficiently compare CRM evolution, the organisms must be divergent enough to have clear genotypic differences but close enough to be able to identify homologous CRMs. (Davidson and de-Leon 2007)

How to create biological diversity has been at the centre of evolutionary biology for over 150 years. Simplistically, this takes place in two steps: firstly by introducing variability

(Olson-Manning et al., 2012) that secondly, over time, by random genetic drift or selection will be fixed in the population. The accumulation of these genotypic alterations gives access to different evolutionary paths.

Many different mutations within the coding part of DNA or within non-coding cis-regulatory regions can have an impact on the GRNs and on development; examples are discussed below. To understand the reasons for the genetic changes, the evolutionary paths have to be mapped to then understand the mechanisms that led to these changes. Changes can be more or less important, ranging from point mutations to insertions or deletions. Eric Davidson argues that the impact of these mutations depends also on the position of the affected gene within the GRNs. Changes within these GRNs can have an effect on development but also on evolution.

3. Regulatory evolution as a tool to drive and change morphogenesis

How can GRN evolution create morphological diversity? Do we find that there is a preferred route to creating this diversity (Prudhomme 2009, Halfon 2017)?

i. The simplest changes are in differentiation gene batteries

The simplest change to implement is on a gene on the periphery of the GRNs: the differentiation gene batteries (Figure 8A). The activity of these genes is at the terminal point of the developmental process and they give little feedback to the internal GRNS (Erwin and Davidson, 2009).

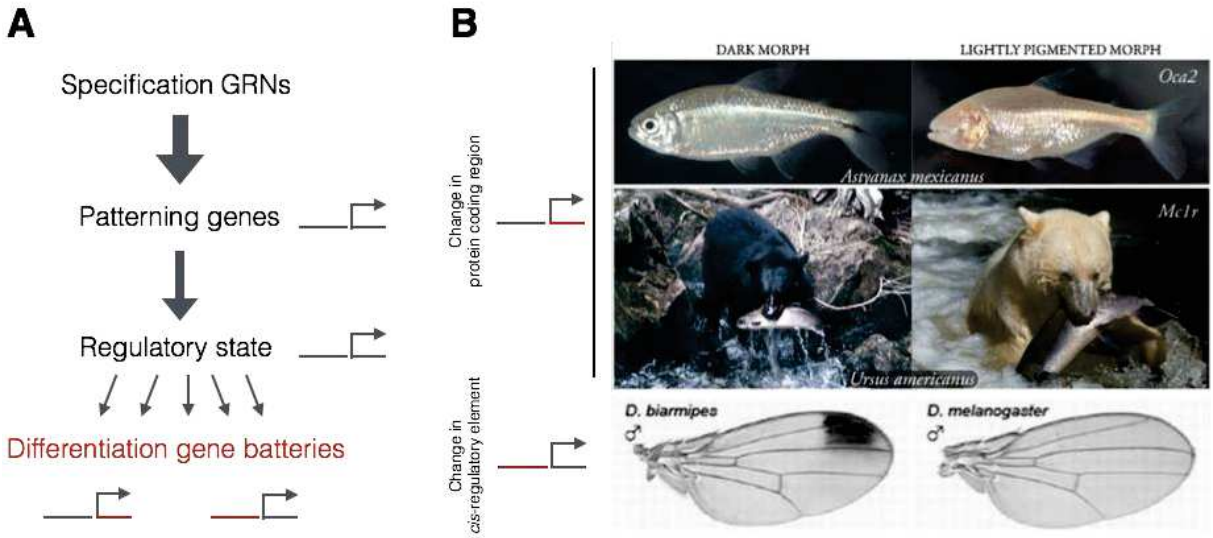


Figure 8. A) Representation of the position of differentiation gene batteries in the GRNs. B) Mutations in the protein coding sequences of genes, Oca2 and Mcr1, and in the cis-regulatory sequence of yellow all affecting pigmentation (adapted from Gaunt and Paul, 2012; Gompel and Prud'homme, 2009).

There are two possible ways to affect these periphery genes; a first is by a mutation within the coding region that will affect protein function. A feature that has commonly been changed across species is pigmentation. There are many genes that can affect pigmentation but very often, changes are found within a same set of genes (Gompel and Prud'homme, 2009). Two such genes are Oca2 gene (involved in the production of

melanin) in the Mexican cavefish *Astyanax mexicanus* (Protas et al., 2006) and in humans (Oetting et al., 2005) and the Mc1r gene (a receptor that controls the type of melanin produced) across many species (Hoekstra and Nachman, 2003; Mundy, 2005; Steiner et al., 2009) (Figure 8B). What is particularly interesting with these examples is that gene size has also played an important role in their repeated evolution. Indeed, a longer coding sequence can increase the chance of mutations that could lead to a functional change (see review Gompel and Prud'homme, 2009).

A second way to affect a differentiation gene battery is by changing its expression pattern (Erwin and Davidson, 2009). An example is *yellow* in *Drosophila,* which is an enzyme necessary for the melanin pigment; *Drosophila biarmipes* has a wing spot that *Drosophila melanogaster* is lacking (Figure 8B). *D. biarmipes* has actually gained binding sites for repressor Engrailed within its *cis*-regulatory sequence (Gompel et al., 2005; Wittkopp et al., 2002).

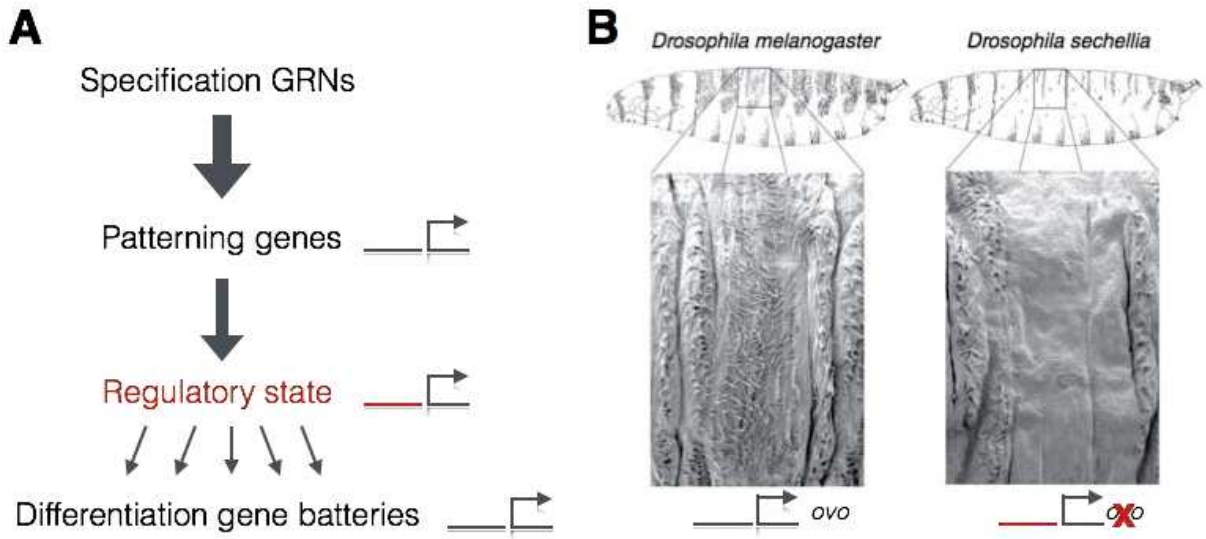


Figure 9. A) Images of trichomes in Drosophila melanogaster which were lost in Drosophila sechellia due to multiple mutations in the cis-regulatory region driving ovo expression (adapted from Stern and Orgogozo, 2009). B) Representation of the position within the GRNs of genes defining the regulatory state.

ii. Regulatory genes driving differentiation gene batteries

To change the expression of a differentiation battery gene, a change can occur in the expression pattern of a regulatory gene that drives it (Figure 9A). An example is the case in *Drosophila sechellia,* which lost the dorsal bands of trichome (Figure 9B); the gene responsible is TF *ovo*, also known as *shavenbaby*, that interprets patterning genes and then drives structural genes. The position of this gene within the GRN accounts for its repeated evolution across species (Stern and Orgogozo, 2008, 2009). The loss of expression of *ovo* in *D. sechellia* is due to a series of multiple nucleotide substitutions in five of its *cis*-regulatory regions, each substitution contributing weakly to the phenotype (Frankel et al., 2010).

iii. Cis-regulatory evolution is favoured for genes positioned internally in the GRNS

Sean Carroll proposes that mutations in *cis*-regulatory sequences are more likely to be selected by evolution than mutations in coding sequences. This is especially true for genes that are found in the internal linkages of developmental GRNs. Due to their position, they affect many tissues; these genes are pleiotropic (Erwin and Davidson, 2009). Because of their pleiotropic nature, and the modular nature of enhancers, a mutation in a *cis*-regulatory sequence will only affect the gene's function in a specific territory, thereby decreasing the chance of lethal or strongly deleterious effects (Carroll, 2008). For example, the vertebrate *Pitx1* transcription factor is required for hindlimb and craniofacial development, and its inactivation in the mouse is lethal (Lanctôt et al., 1999; Marcil et al., 2003; Szeto et al., 1999). Yet, mutations in the Pitx1 locus is causal in several cases of adaptive reduction of the pelvic fin (the equivalent of tetrapod hindlimbs) in freshwater stickleback fish (Figure 10). In all cases, the mutations are found in a *cis*-regulatory module driving specific expression of Pitx1 in the pelvic fin precursors (Shapiro et al., 2004). Interestingly, the mutated enhancer is located next to a site of lagging during DNA replication making this region more prone to DNA breakage (data presented by David Kingsley).

Changes in the coding regions of a gene will have a more immediate phenotypic effect. They can also occur more easily than *cis*-regulatory changes that need a collection of mutations to have an effect (Landry et al., 2007), however, *cis*-regulatory changes will be retained over time more readily due to the fitness cost of changes in the coding gene (Otto, 2004).

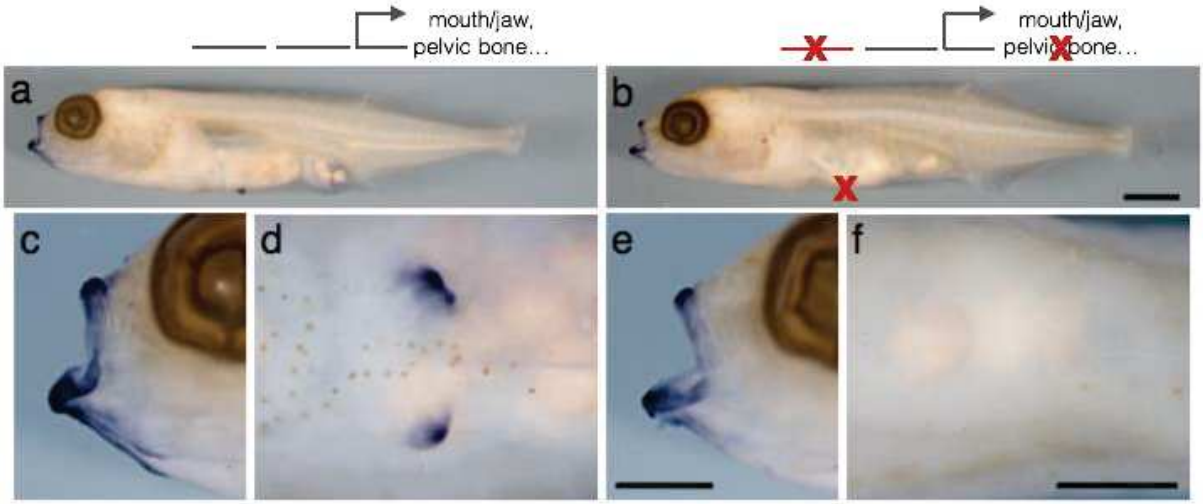


Figure 10. Images of Pitx1 expression in the stickleback fish: (a,c and d) show expression in the mouth and pelvic bone and (b, e and f) show expression only in the mouth due to the loss of the enhancer driving Pitx1 expression in the pelvic bone (adapted from Shapiro et al., 2006).

iv. Same morphology, different network: convergence and DSD

Of the examples previous mentioned that affect morphology, a few involved genes controlling pigmentation, such as Mcr1 and Oca2. In fact, there are over 100 genes

controlling skin or fur colour (Bennett and Lamoreux, 2003). This high number of genes contributing to a same phenotype increases the chance that a change in any of these genes will produce the same phenotype. Indeed, skin colour is great example of how several different evolutionary changes can arrive at the same result separately; this is known as convergence.

There also exist cases where change in the GRNs has not changed the outcome; this is known as developmental systems drift (DSD) or phenogenetic drift (True and Haag, 2001). Literature about GRN evolution usually disregards DSD despite the fact that it is a particularly interesting form of GRN evolution. An example of DSD can be found in the development of mosquito *Aedes aegypti* nervous system that appears identical to that in *Drosophila* but A. aegypti has co-opted the GRN driving ventral midline development in *Drosophila* and shifted it to later embryonic development without any apparent differences in overall nervous system morphologies (Suryamohan et al., 2016). Further CRM analysis is necessary to understand the mechanisms to this GRN co-option (Halfon, 2017).

4. Theories behind GRN evolution

An accumulation of information on regulatory evolution has allowed evolutionary biologists to develop several, sometimes conflicting, theories about the control of morphological change by GRNs. Below, some of these mechanisms to circumvent the complex structure of GRNs to still produce diversity are discussed.

i. Is the structure of GRNs defining the rate of their evolution?

The development of animal body plan is driven by GRNs; therefore, evolutionary changes affecting the body plan generally originate from changes in the GRNs as exemplified for dog morphologies (Parker et al., 2009; Sutter et al., 2007). Eric Davidson and Douglas Erwin proposed 10 years ago that GRNs do not evolve homogeneously. Rather, they are made up of several layers of structures differing in their ability to change in the course of evolution. At the start of the GRNs, the networks are responsible for the tissue specification; next, they organise pattern formation for morphological structures. At the very end of the networks, there are the differentiation gene batteries that confer their differentiated features to the tissue or body part (Erwin and Davidson, 2009). These layers are represented in a schematic (Figure 11). This theory proposes why there has been so little change in body plans since the early Cambrian period yet on-going occurrences of novel features characterising speciation (Davidson and Erwin, 2006; Erwin and Davidson, 2009).

Davidson describes the hierarchical levels of the GRNs as follows: the complex and therefore evolutionarily inflexible “kernels” that control body parts (mostly transcription factors), small “plug-ins” that reoccur in many of the developmental processes (mostly involved in signalling), input/output switches “I/O” that function as switches turning on or off a certain sub-network and finally the differentiation gene batteries at the end of the networks. The function of the kernels and the “plug-ins” is purely regulatory. Differentiation gene batteries are found at the periphery of the GRNs as they are the final product of the networks. These differentiation gene batteries are amenable to evolution and allow “renovation” (Erwin and Davidson, 2009).

At the core of developmental GRNs, kernels are proposed by Davidson to be so complex in their architecture that they are unable to change. Davidson and Erwin argue that all of the TFs in combination are required for the *cis*-regulatory elements to be functional and therefore if one of the genes of the kernel is not expressed, then it has a lethal phenotype. This recursive wiring and their role in controlling body plan prohibit kernels from change. They propose that the structure of GRNs means that once body parts defined, they cannot be changed.

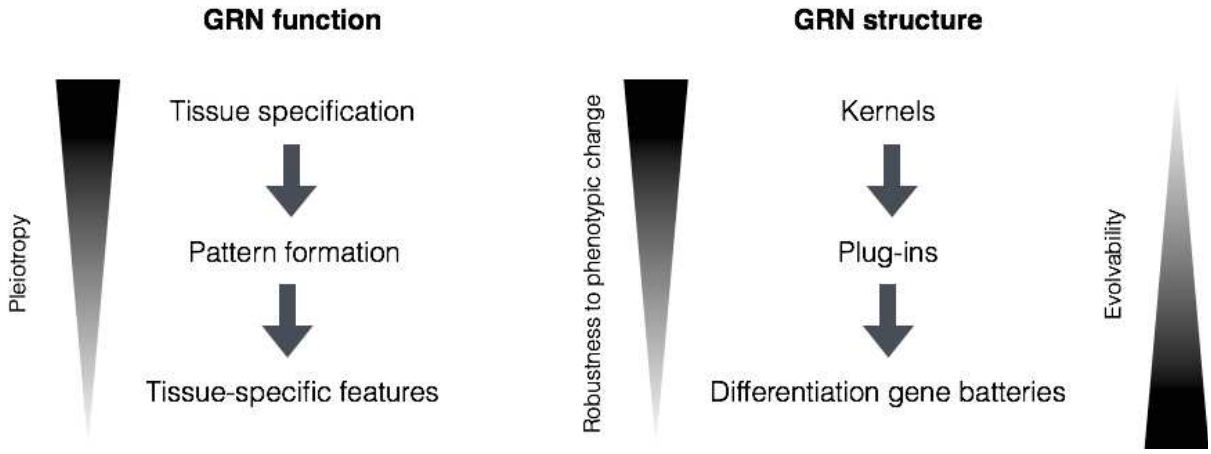


Figure 11. A schematic of the structure of the GRNs for each developmental step demonstrating the pleiotropic nature of each layer of the networks; the more pleiotropic the GRNs, the more robust they are to change and evolvability.

"Plug-ins" are also conserved as they are involved in many earlier processes. These are for example signal transduction systems such as Wnt and Notch signalling (Artavanis-Tsakonas et al., 1995; Cadigan and Nusse, 1997). Although these networks are conserved, their uses may not be; they can be involved in patterning activity of small circuits in a myriad of different regions of the organism (Davidson and Erwin, 2006).

Since the Cambrian period, changes in animal development are dependent on the 3 classes of GRNs which each have different consequences and different rates of evolution. Therefore, Davidson and Erwin argue that microevolution based on neutral mutation at a single base or gene duplications that occur in a temporally stable manner cannot satisfactorily explain animal developmental evolution. However, this would mean that in networks so unchangeable, the only part of the GRNs that are evolvable would be at the periphery (Davidson and Erwin, 2006).

ii. The role of neutral mutations in evolution

The "kernel" hypothesis, which posits that robustness leads to evolutionary stasis, is challenged by the work of Andreas Wagner, who argues instead that robustness can actually promote evolution.

Robustness of a system is defined as its resistance, without negative or positive effect on its fitness, to perturbations induced by mutations or environmental changes. The more robust a system is, the more phenotypically neutral mutations it can accumulate. Robustness to mutations of environmental changes thus buffers the phenotypic effect of these changes, and so should intuitively hinder the evolvability of the system. However, these new mutations also give access to novel regions of the genotypic space, that is, to novel and potentially non-neutral mutations. Here ensues Andreas Wagner's conundrum: does robustness promote or hinder evolvability (Wagner, 2005, 2012)? On one hand, neutral mutations cannot be selected for through natural selection. However, on the other hand, even though this neutral mutation will not affect the main functions of the organism, it will, however, lay the foundations for future innovations.

The theory of neutral genetic change, first introduced by Kimura in 1968, confronted the general opinion at the time that no mutation was neutral. It was thought that due to large population sizes in mammals, for example, if a neutral mutation were to occur, it would not be able to stick and would therefore be lost. Kimura noted, however, that there were several works showing surprisingly high rates of mutations such as in haemoglobin amino acid sequences across mammals. He insisted that neutral mutations are a common occurrence and concluded that genetic drift is a valid evolutionary mechanism (Kimura, 1968).

Neutral mutations are actually abundant in robust systems however they do not remain neutral. It is the accumulation of neutral mutations that promotes evolvability: these mutations allow the species to explore a larger domain of genotypic space (Ciliberti et al., 2007; Wagner, 2005). A collection of neutral mutations could perhaps contribute to changing the architecture of an enhancer. First mutations could remain neutral at the phenotypic level or even at the GRN level if one TFBS out of many is created or lost. Potentially, this additional TFBS could lead to a slight increase in enhancer activity or gene expression level. However, the additional site could have an insignificant effect on CRM activity. A novel, neutral TFBS site could come to use if another TFBS for the same TF is lost through mutations: this would result in TFBS turnover.

A neutral mutation could be neutral for one aspect of development whilst leading to a slight change in phenotype in another (A Wagner 2005). In this case, the change has had a pleiotropic effect, which, through compensation, could promote further change.

iii. Pleiotropy drives evolution

It was discussed above that *cis*-regulatory mutations were preferred in evolution because they are less pleiotropic than genes (Otto, 2004). However, Pavlicev and Wagner argue that mutations in pleiotropic genes may still be selected during evolution. In their paper, they proposed the Selection-Pleiotropy-Compensation model based on universal pleiotropy that states that most genes affect different characteristics. The model proposes that if there is a fixed mutation in a pleiotropic gene due to its positive effects on fitness outweighing the negatives, then a second compensatory change will be selected to compensate for the negative effects (Pavlicev and Wagner, 2012). Perhaps the adaptation and compensation mechanisms contribute to the complexity of the network, which in time, could lead to the formation of kernels.

In the case of divergence in homologous body parts, such as limbs in humans and bats (Young and Hallgrímsson, 2005; Young et al., 2010), diverged from ancestrally shared developmental programs. In such cases, as a rule, the more divergent the feature is, the less pleiotropic it becomes. Young *et al.* propose that decreased pleiotropy is achieved through increased inhibition (Young et al., 2015). This feature seems to be what Davidson and Erwin called the "plug-ins" (Erwin and Davidson, 2009). Although they argued that plug-ins are not amenable to change, their pleiotropic nature could leave them vulnerable to adaptive compensation and could make them a source of change.

III. Ascidians as a model organism

Compensation at the molecular level can be seen quite extensively in regulatory sequences, manifested by TFBS turnover (He et al., 2011). Ascidians are particularly fascinating model organisms to study this evolutionary phenomenon. Despite their fast evolving genomes, they have conserved an almost identical development with stereotyped cell lineages.

1. An introduction to ascidian history

Ascidians have a long history in biology due to their unique appearance making their phylogeny long debated until recently.

i. Ascidian phylogeny

Ascidians, also known as sea squirts, are marine invertebrate chordates that can be found in both shallow and deep waters around the world. These sessile animals fix themselves to a surface, such as a rock, and filter seawater for plankton through their two siphons. The origin of their name is Greek "askidion" meaning a little wineskin with two holes reflects their appearance. Their body is covered by a cellulose tunic, from which the subphylum Tunicata gets its name; tunicates are the only animals to be able to synthesise cellulose.

Originally, Linnaeus classified ascidians as mollusks in 1789, due to their soft bodies; however, in 1816 Lamarck doubted this classification because their body organisations were too specific. Later in 1886, the Russian embryologist Kowalevsky noted that the tadpole-like ascidian and amphioxus larvae had a central notochord (Figure 12A) and merged these two taxa into a new group, the protochordates, which he placed between vertebrates and invertebrates, thinking at the time that he had bridged the evolutionary gap (Kowalevsky, 1886).

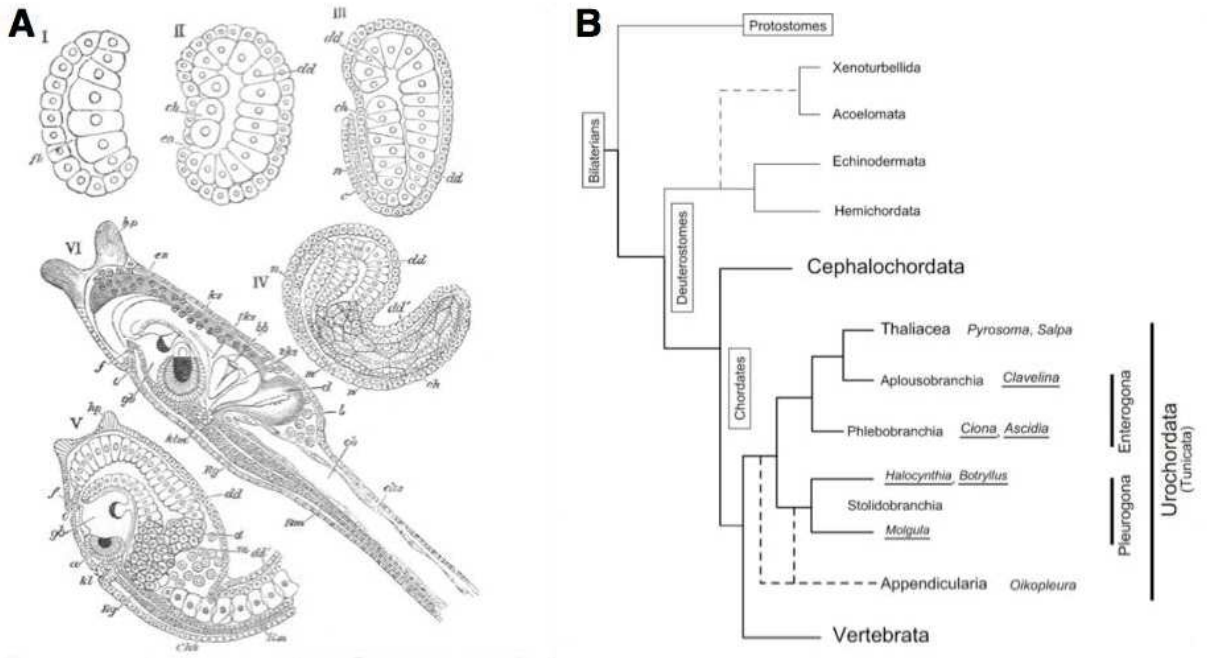


Figure 12. Deciphering the origin of ascidians. A) Schematics of ascidian embryos by Kowalevsky (Kowalevsky, 1886). B) The phylogenetic tree placing tunicates as a sister group of vertebrates (Satoh, 2014).

Following genome sequencing of *Ciona intestinalis* (Dehal et al., 2002a), molecular phylogeny approaches revealed the tunicates' position as a sister group of vertebrates in 2006 (Figure 12B) (Delsuc et al., 2006). Ascidians are part of the class Ascidiacea, subphylum Tunicata and phylum Chordata. Although the precise phylogenetic relations within tunicates are not fully elucidated, they are split between three groups: thalacians, ascidians and larvaceans (Lemaire, 2011). Ascidians however may not constitute a monophyletic group (Tsagkogeorga et al., 2009).

Ascidians include both solitary species, which reproduce sexually only, and colonial species, which reproduce both sexually and by budding. The most studied solitary ascidian is *Ciona intestinalis* which was recently shown to form a complex of two species at least (Pennati et al., 2015). The most widely used of the two is now named *Ciona robusta*, based on Hoshinoa and Tokioka's 1967 morphology description (Hoshino and Tokioka, 1967) and the other, which can be found throughout the Northern Atlantic including around the coasts of France and has retained the name *Ciona intestinalis*.

ii. Genome features

The *Ciona robusta* genome was the first ascidian genome to be sequenced, in 2002. Its genome is highly polymorphic and compacted; it has a size of 160Mb of which 117Mb is euchromatic. Within these 117Mb, 15,250 protein-coding genes were predicted meaning that there is one gene/ 7,7kb on average, which is just slightly higher than *Drosophila* at one gene/ 8,9kb. A similar density is found in the *Ciona savignyi* genome (Small et al., 2007). The *Ciona robusta* genome is particularly AT rich (65%) compared to the human genome (45%) (Denoeud et al., 2010). Furthermore, ascidians mostly have single copy genes, in contrast to the many paralogs found in vertebrates (Dehal et al., 2002b).

Within *Ciona species*, the genomes have undergone multiple rearrangement; the syntheny is limited to sections under 1Mb in length between *Ciona robusta* and *Ciona savignyi* (Hill et al., 2008). Furthermore, *Ciona robusta* and *Ciona savignyi* have a high genomic diversity: 1-2% and 3-5% polymorphism respectively (Abdul-Wajid et al., 2014). This polymorphism can facilitate micro-evolutionary studies.

2. Ascidian development

The ascidian development is quite remarkable. Its rapid development, small cell number and and stereotyped cell lineages make them excellent model organisms.

i. Stereotyped embryogenesis

The first known experimental work on ascidian embryos was by Chabry in 1887 on *Ascidiella aspersa*; in part of his work, Chabry destroyed half of the embryo at the 2-cell stage and witnessed that the embryo developed into a half-larva, rather than a dwarf larva. In 1905, Edwin Conklin first described the invariant ascidian cell lineages and the segregation of a coloured posterior vegetal cytoplasmic region of the egg, the yellow crescent, into larval muscle suggesting that the precise inheritance during cell division of localized maternal factors governs cell differentiation. This was confirmed for the muscle lineage in 1973 by Whittaker in cleavage-arrested embryos, but shown to be inaccurate for most cell fates, whose development, as in vertebrates, relies on cell-cell communication.

An adult solitary ascidian has two openings: an oral siphon that acts as the mouth and an outcurrent atrial siphon. Ascidians are hermaphrodites; their reproduction happens at sunrise at which time they release their gametes into the sea. In the lab, the eggs and sperm are collected by dissection. These produce rapidly developing stereotypic embryogenesis, bilaterally symmetric cleavages (Figure 13). Enclosed in a chorion, *Ciona robusta* eggs are roughly 140µm in diameter. Even before fertilisation, the egg has already defined the animal and vegetal pole. After fertilisation, the cytoplasm is reorganised to localise the maternal factors (Sardet et al., 2007). The embryo then undergoes a series of stereotyped divisions that are very well conserved. For example, distantly related *Ciona* and *Phallusia* have an almost identical development.

At 18°C, gastrulation starts at 5hpf when the embryo only counts 112 cells; this is followed by neurulation. A tailbud is formed to eventually produce a tadpole-like larva at 18hpf. The ascidian tadpole has distinct tissues (including epidermis, endoderm, notochord, muscle, mesenchyme and central and peripheral nervous system) even though it is only roughly 2600 cells. The larva will swim around until it fixed to a surface; it will go through metamorphosis into a sessile juvenile which starts feeding.

Thanks to the invariant cleavage pattern of ascidian embryos, individual embryonic cells can each be individually named and found across embryos. The nomenclature adopted, {a,b, A, B}.p.q (e. g. A4.1, b5.2 or B6.7) is based on Conklin's scheme (Conklin, 1905). Letters "a" and "b" denote the progeny of the anterior and posterior animal cells at the 8-cell stage respectively, while "A" and "B" are the progeny of the anterior and posterior

vegetal cells. The first number, p, is the rank of the somatic cell cycle since fertilization (e. g. 5 for a 16-cell stage embryo) and the second number, q, is a personal cell identifier within the lineage: cell q will upon division give rise to daughters 2q and 2q-1, the latter being closest to the vegetal pole. Systematic tracking of the larval progeny of early pregastrula blastomeres (Nishida, 1987) revealed that by 112-cell stage, almost all cells will contribute to a single tissue, a remarkably early fate restriction.

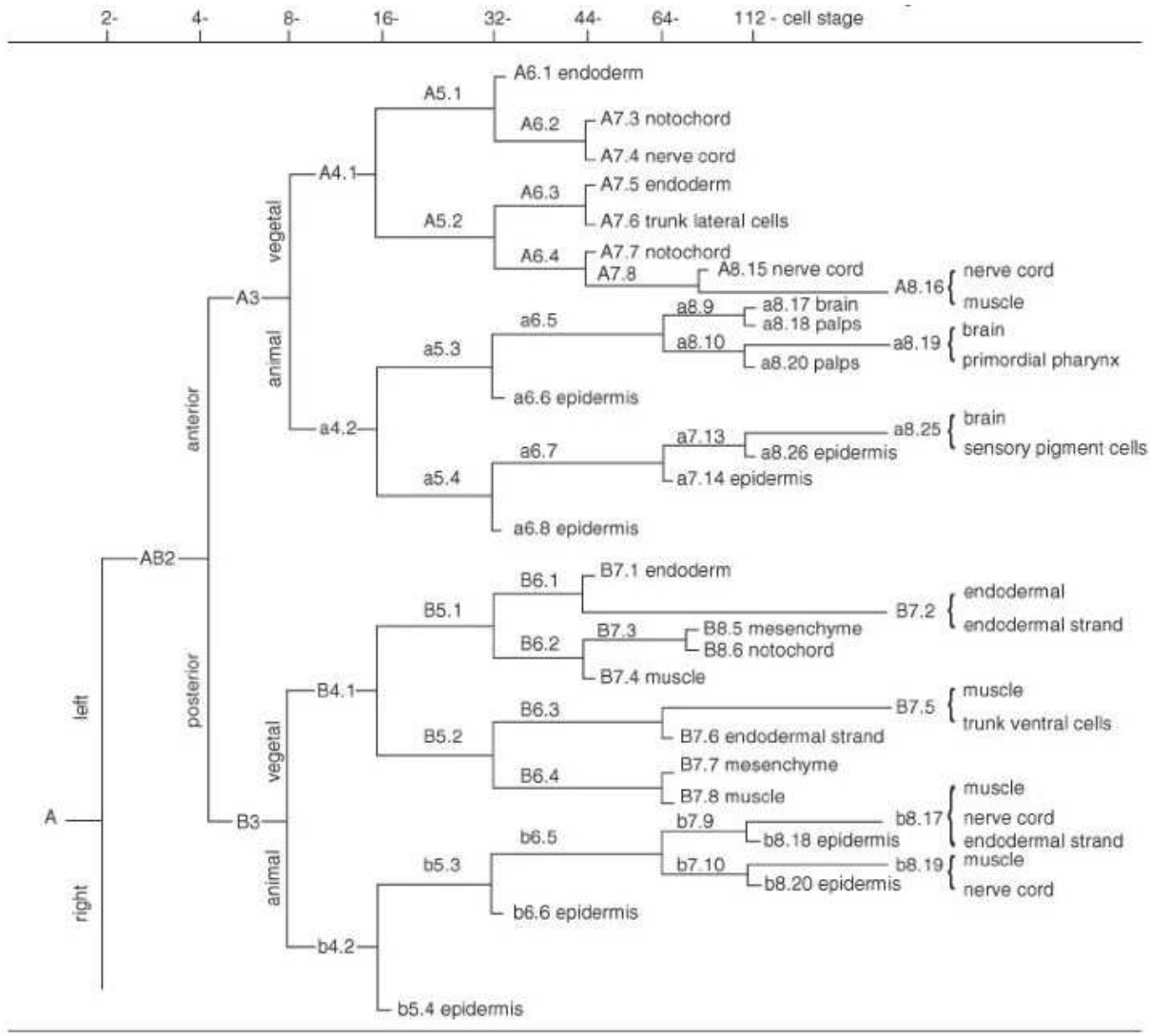


Figure 13. Ciona cell lineages up to 112-cell stage and tissues fates in the ascidian embryo; it is bilateral so only one half of the embryo has been represented (Satoh, 2014).

A rich palette of molecular techniques have been adapted to ascidians such as: *in situ* hybridisation (Christiaen et al., 2009a), treatment with pharmaceutical inhibitors, microinjection of morpholinos antisense oligonucleotides (Satou et al., 2001a), CRISPR-Cas9-mediated knock out(Christiaen et al., 2009b, 2009c; Stolfi et al., 2014; Treen et al., 2014) and, the most popular, electroporation of expression constructs into fertilized

eggs (Corbo et al., 1997). These techniques combined with a rapid, stereotyped development and low cell number, have turned ascidians into a great model system to study GRNs and morphogenesis with a cellular level of resolution. Several reviews summarize the work of the ascidian community in this field (Hudson, 2016; Kubo et al., 2009; Lemaire, 2009), we will focus here on what is known about the formation and function ascidian embryonic endoderm.

ii. Ascidian endoderm

From the first stages of ascidian embryogenesis, the endoderm is a central tissue. It drives tissue fate specification by signalling neighbouring cells to adopt mesodermal. In parallel, it undergoes dynamic cell shape changes which drive its invagination during the first phase of gastrulation. The endodermal tissue is a particularly easy tissue to work on in the earlier stages as the cells are large and maintain a central position. Furthermore, leading up to the initiation of gastrulation, the endodermal precursors undergo few rounds of cell division meaning that at the onset of gastrulation there are only 10 endodermal precursors.

The endodermal cell lineages have been well characterised; originating mainly from the A- and B-line but also in part from the cells descending from b8.17 and perhaps B7.6 (Figure 14A). As early as the 32-cell stage, the vegetal A6.1 and B6.1 cell pairs are already restricted to the endodermal cell fate (Figure 14B); by the 64-cell stage, anterior vegetal cell pairs A7.1, A7.2, A7.5 and posterior vegetal cell pairs B7.1 and B7.2 are restricted to the endodermal fate. In the Japanese solitary ascidian *Halocynthia roretzi*, these 10 endodermal cells will divide only 5 or 6 times before the larval stage to give about 500 cells in total (Hirano and Nishida, 2000). It is likely that the endoderm includes a similar number of cells in *Ciona* . During metamorphosis, these endodermal cells develop into digestive organs including the oesophagus, stomach, intestine, branchial sac and endostyle.

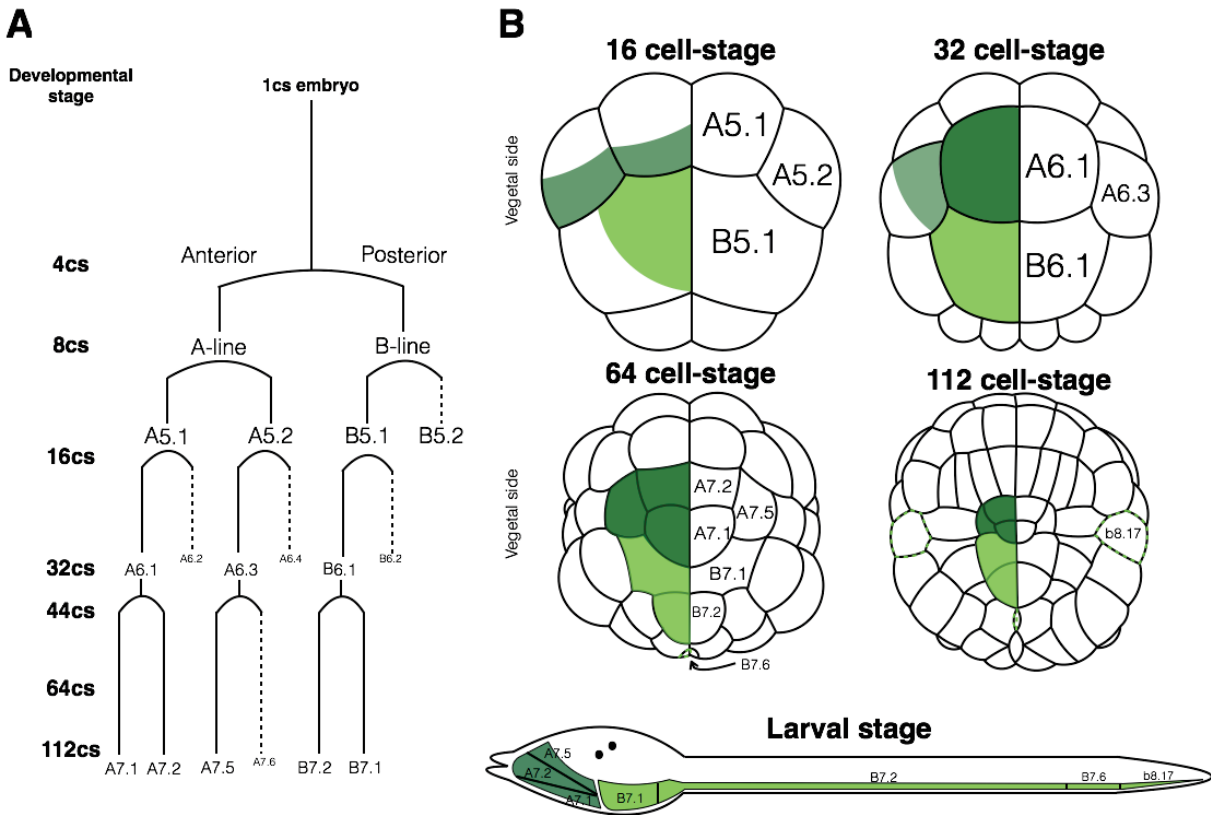


Figure 14. Ascidian endoderm throughout development. A) Major A- and B-line endodermal cell lineages up to 112-cell stage. B) Schematics of Phallusia mammillata from 16-cell stage up to hatching larva. The naming at the larval stage represents where the cells at these locations originate from. Dark green is A-line endoderm, light green is B-line endoderm. The cells named on the schematic will contribute to the endoderm (adapted from Satoh, 2014).

3. Endoderm GRNs

Historically, the best-examined tissue in ascidians has been the notochord. It is a defining feature of chordates and therefore decoding the GRNs that drive its development in ascidians could be a stepping-stone towards elucidating notochord evolution within chordates or even chordate evolution. The endoderm, however, is a particularly exciting tissue; in a few short hours, it is responsible not only for coordinating the body plan by inducing the mesodermal tissue but also for driving the first major morphogenetic event, gastrulation.

i. Building ascidian GRNs

Recent advances in identifying regulatory genes involved in developmental processes have gone a long way for building GRNs in ascidians. Constructing the GRNs can help elucidate how genes coordinate the development of the ascidian body plan. Ascidians are excellent models for studying GRNs at the cellular level. They have few cells, tissue fate specification occurs very early in development and they have few regulatory genes

(Imai 2006). The Satou lab has already done extensive work on the *Ciona* GRNs; this work is provided on their Ghost database (Satou et al., 2005).

To start deciphering the ascidian GRNs, it was first necessary to have a precise detail of wild type gene expression at stages close together. A lot of this work was performed in *Ciona robusta* by Imai, in Satou's lab at the time. *Ciona robusta* has about 318 transcription factor genes and 110 genes encoding major signalling molecules all together; based on WISH assays, there are only 53 TF genes and 23 SPM genes zygotically expressed between 16 cell-stage and early gastrula stage (Imai et al., 2006).

The Ghost database incorporates expression data and gene regulatory information to build *Ciona* GRNs at the cellular level from egg up to mid tailbud (Figure 15A) (Satou et al., 2005). The information provided is based on expression data from wild type embryos and knockdown experiments in *Ciona robusta* and *Ciona savygnyi* heavily based on the 2006 work of Imai *et al.* but also 28 other papers. Last updated in 2014, these networks incorporate information at the cellular level about 112 regulatory genes. The genes have been represented in alphabetical order with the backdrop of the whole embryo summary. The endodermal GRNs can be navigated more easily when they are represented in order of developmental stages (Figure 15B).

The Ghost GRNs provide a highly informative view of the regulatory activity of studied *Ciona* genes; however, this representation lacks the perspective to judge the direct impact of the GRNs on fate specification or morphogenesis. By integrating knockdown phenotype information to the representation of the networks, it would be possible to begin to grasp the importance of certain networks driving specific *Ciona* developmental processes.

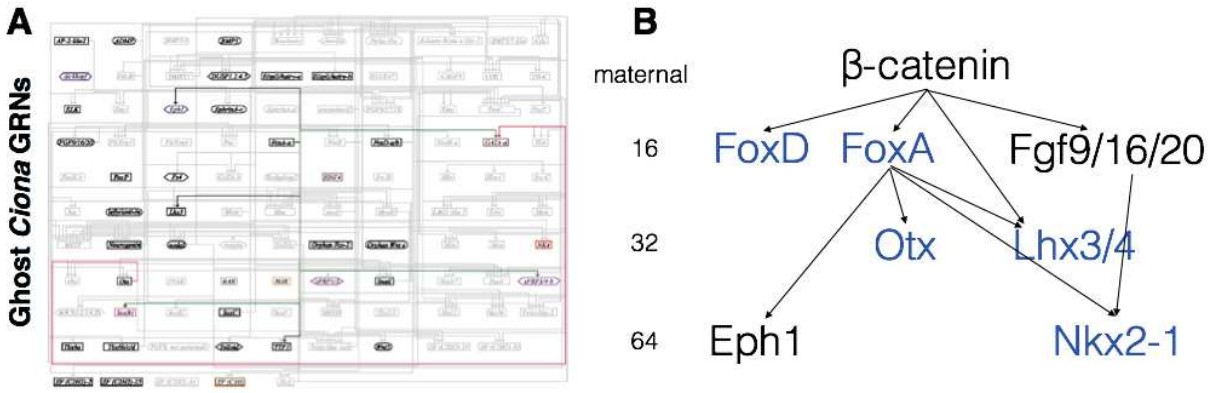


Figure 15. Endodermal GRNs A) Summary of the GRNs in the whole endoderm from egg up to mid tailbud with a backdrop of the summary of the whole embryo GRNs (Satou et al., 2005). B) Ghost Ciona endodermal GRNs up to early gastrula (adapted from Ghost GRNs).

ii. Endodermal cell fate specification events

Maternal β-catenin is known to be the initiator of the specification of endodermal fate in many metazoans, including ascidians. Early experiments in *Halocynthia roretzi* involving egg cytoplasm transplantation showed that maternal factors located in the vegetal

cortical cytoplasm are responsible for the autonomous differentiation of endoderm cells (Nishida, 1996). Mis-expressed or constitutively active of β-catenin causes mis- or over-expression of the endodermal marker alkaline phosphatase (Figure 16A) (Imai et al., 2000). Alkaline phosphatase (AP) is located in the endoderm at the larval stage in *Ciona* and *Halocynthia* although it does not seem to be expressed in *Phallusia*.

The nuclear accumulation of maternal β-catenin in the mesendodermal cells is the first step towards cell fate specification. Although maternal ß-catenin is uniformly distributed throughout the cytoplasm in the first stages of embryo development, it is found in the nucleus mainly in the vegetal hemisphere, from the 16-cell stage onwards. Once nucleated, β-catenin acts with TCF to activate the first zygotically expressed genes, some of which are involved in transcription and signalling. These include FoxA, FoxD and Fgf9/16/20, which start the cascade of expression of genes driving the first specification events in the mesendoderm cells (Hudson et al., 2016), such as Lhx3/4, necessary for the induction of the endoderm fate, and Nkx2-1, that contributes to the endodermal fate (Ristoratore et al., 1999; Satou et al., 2001). FoxD has been found to act as both an activator and a repressor; it promotes endomesodermal fates but also inhibits ectodermal fate in the vegetal hemisphere. RNA-seq at the 32- and 64-cell stages in FoxD morpholino *Ciona* embryos that FoxD activates Zic-r.b and Lhx3/4 (Tokuhiro et al., 2017).

There are distinct mechanisms that define the endoderm and mesoderm fates in ascidians. A first is the activation of β-catenin yet again, at the 32-cell stage, but this time only in the endoderm progenitor cells. If this 2nd wave of β-catenin activation is suppressed, for instance by microinjection of dominant negative TCF (TCFΔC) into an endodermal cell, endoderm genes such as Lhx3/4 and Nkx2-1 are suppressed and the mesoderm marker gene Zic-r.b is expressed instead (Figure 16B). This series of binary β-catenin decisions has been termed ON-OFF for mesoderm and ON-ON for the endoderm (Hudson et al., 2013).

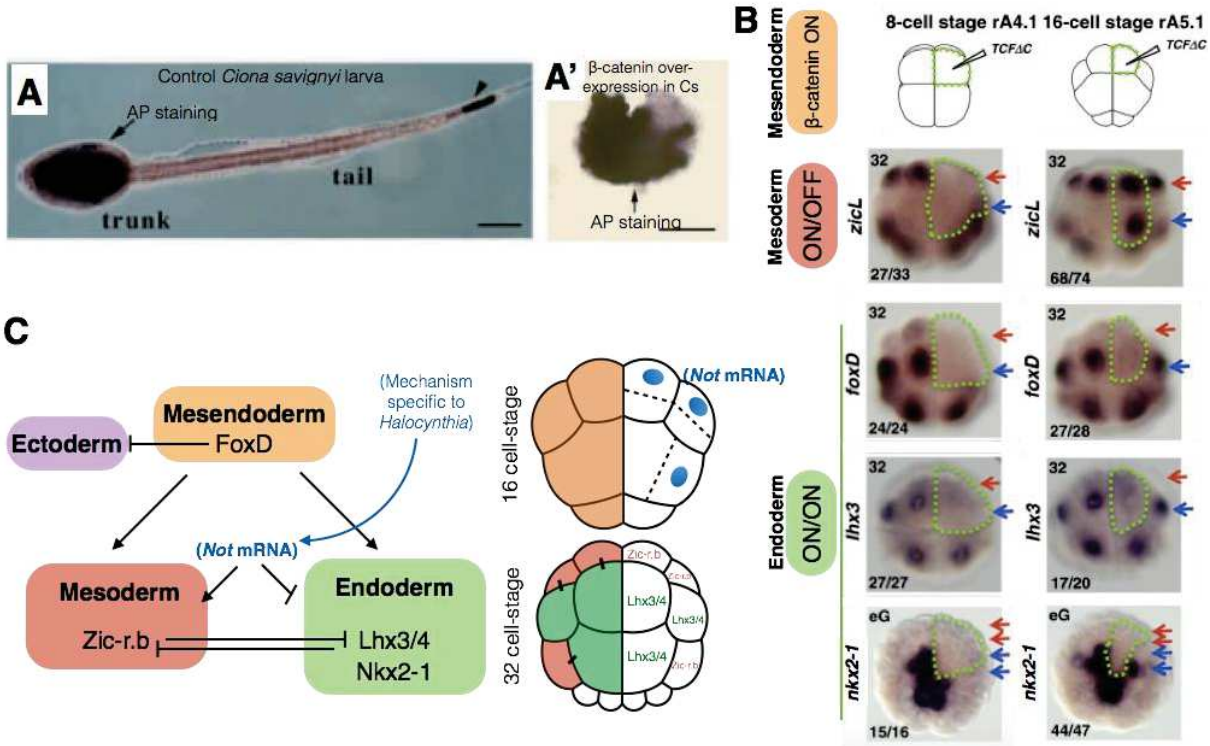


Figure 16. The early drivers of the endodermal fate. A) Histochemical staining of AP endoderm marker in Ciona savignyi larvae; (A) Embryo injected with LacZ mRNA and (A') embryo injected with Cs-β-catenin mRNA. B) In situ hybridisation images of Ciona intestinalis embryos injected with TCFΔC in the A4.1 cell at the 8-cell stage compared to embryos injected in the A5.1 cell at the 16-cell stage. The green dotted circle highlights the daughter cells of the injected cell; the red arrow should be the mesodermal cell and blue should be an endodermal cell (Hudson et al., 2013). C) The endoderm regulatory network in Halocynthia roretzi (Adapted from Takatori et al., 2010).

Lhx3/4 is sufficient to suppress the mesodermal fate and the Zic-r.b, expressed in the mesoderm, suppresses the endodermal fate. In *Halocynthia,* The segregation of the two cell fates is due to asymmetrical migration of *Not* mRNA to the mesodermal region before cell division (Figure 16C). Therefore, β-catenin alone gives rise to the endoderm whereas β-catenin and Not gives mesoderm (Takatori et al., 2010). This segregation of cell fates coordinated by *Not* is specific to *Halocynthia*, which is a great example of developmental systems drift of a regulatory mechanism in early ascidian development (Hudson et al., 2013).

This opens the question: how conserved are these early endodermal GRNs?

iii. Morphogenetic events driven by the endoderm: gastrulation

Gastrulation is an important morphogenetic step in animal development during which the germ layers of the blastula are rearranged to form the gastrula. The endodermal and mesodermal cells move inwards by invagination as the ectoderm spreads towards the vegetal pole, eventually enveloping the two other germ layers.

In ascidians, gastrulation is a two-step process driven by invagination of the endoderm beginning from the 64-cell stage. The first step is apical-constriction of the 10 endodermal cells following the contractility of the apical actomyosin cytoskeletal network. The first step leads to the endodermal cells adopting a wedge-like shape necessary to complete invagination (Figure 17A). During the second step, tight apices are maintained as apico-basal shortening occurs (Sherrard et al., 2010).

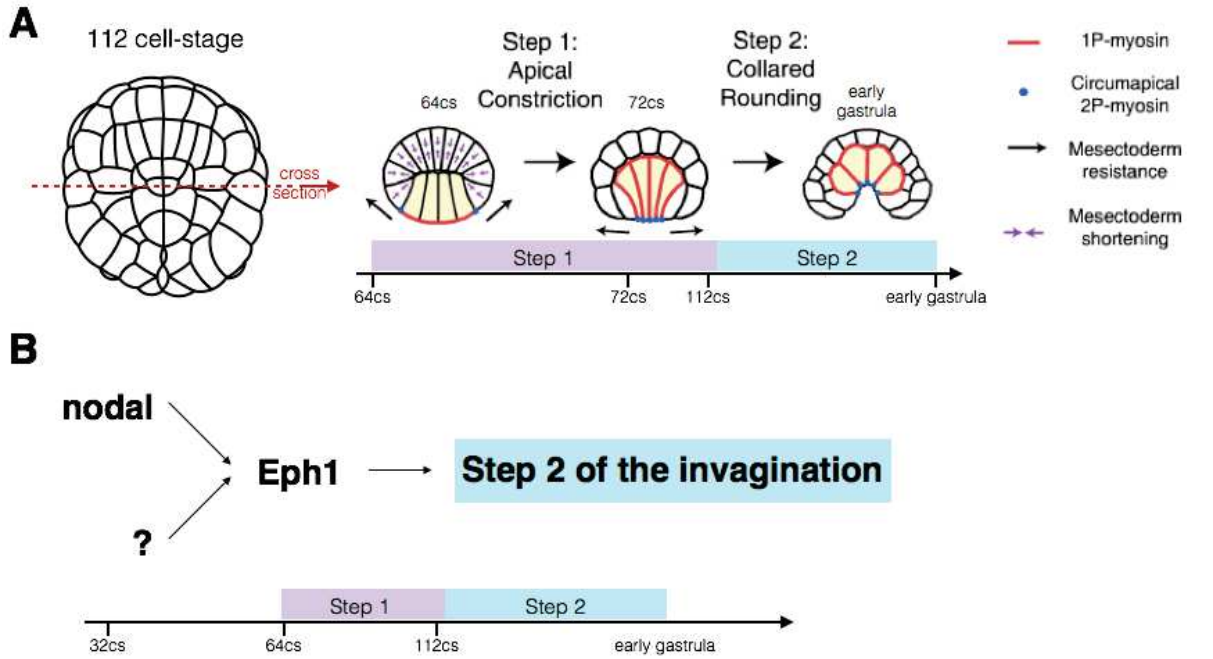


Figure 17. Gastrulation event in Ciona and Phallusia. A) The two-step model of ascidian invagination showing a cross section of an embryo from 64 cell-stage to early gastrula. The endodermal cells are in yellow, step 1 in purple showing the 1P-myosin enrichement in red and the 2P-myosin in blue. Step 2 is in light blue (adapted from Sherrard et al., 2010). B) Regulatory genes driving gastrulation (Based on personal communication, UM Fiùza).

Unpublished data in *Phallusia mammillata* and *Ciona robusta* from the Lemaire and Yasuo labs have elucidated part of regulatory networks that are involved in apico-basal shortening, the second step of the invagination process (personal communication, UM Fiuza). Through morpholino knockdown experiments of Eph1 signalling and over-expression of a dominant negative Nodal receptor or pharmacological inhibition of Nodal signalling, the basolateral relocalisation of the myosin was inhibited. Furthermore, knockdown of nodal signalling decreases expression of Eph1 indicating that nodal is one of the upstream regulators of Eph1 (Figure 17B). Interestingly, although blocking Rho kinase, Nodal or Eph inhibit gastrulation, this did not alter the expresion of both early and late endoderm cell fate markers. Thus the regulatory networks controlling endodermal cell fate specification and gastrulation are in part uncoupled, unlike in vertebrates (Schier and Talbot, 2005; personal communication, UM Fiuza).

4. Regulation in ascidians

So far, we have established that a lot is already known about the early endodermal development in ascidians. However, many of these regulatory interactions still need to be qualified as direct or indirect. One way to confirm the link between the genes is to study the enhancers, and its TFBSs, driving the gene expression in the endoderm.

i. Ascidian regulation: known regulatory sequences

The first ascidian regulatory sequences were discovered in the 90s; the first in *Halocynthia roretzi* driving MA4 expression in the muscle, tested by microinjection (Hikosaka et al., 1994). Corbo *et al.* put in place electroporation in ascidians and found a minimal enhancer driving Brachyury (Corbo et al., 1997). This new technique in ascidians opened the gates to a quicker way to test many regulatory sequences.

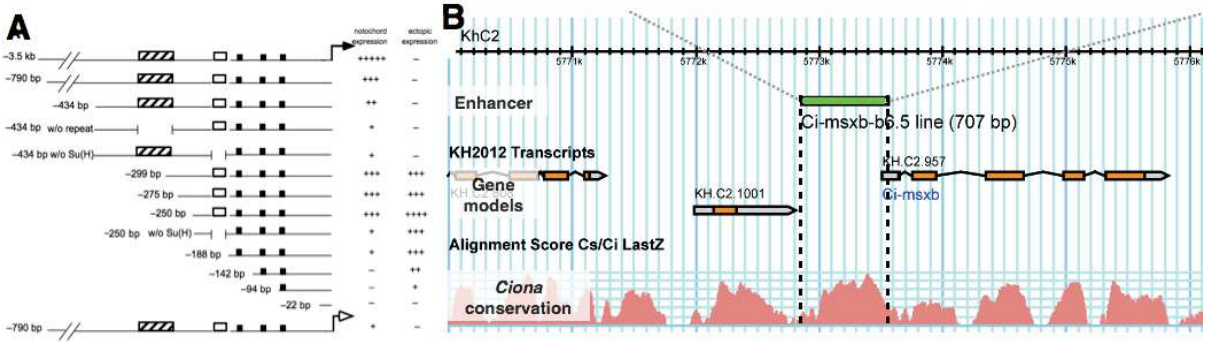


Figure 18. Examples of some methods to identify enhancers in ascidians. A) Serial deletions to regions containing enhancers (Corbo et al., 1997). B) Msxb enhancer is well conserved across Ciona (Roure et al., 2014).

Several methods have been used to find these regulatory sequences in ascidians:

- A common method has been to clone 3kb upstream of the TSS and to successively reduce this sequence by a series of deletions (Corbo et al., 1997) (Figure 18A).
- Some studies randomly chose sequences; this had a low success rate. Harafuji *et al.* cloned sequences roughly 1,7kb on average, 11 enhancers were found out of 138 tested sequences (Harafuji et al., 2002). Using this method, Keys *et al.* had similar success and found 22 enhancers out of 222 tested fragments (Keys et al., 2005).
- Another method is based on conservation of non-coding sequences between distantly related species, such as *Ciona robusta* and *Ciona savignyi*. This was done for *Pitx* (Christiaen et al., 2005). Msxb enhancer is also well conserved across *Ciona* despite changes in the TFBSs (Figure 18B). Interestingly, the latter sequence has maintained the same qualitative activity across species (Roure et al., 2014) which coincides with the flexible billboard model.
- They can also be found by looking for clusters of binding sites for co-expressed TFs (Haeussler et al., 2010).

These different techniques can be combined to yield a higher chance of finding a regulatory sequence.

Most ascidian enhancers have been located within 1,5kb of the TSS. Known minimal ascidian enhancers are less than 200 bp long, much shorter than in other organisms. The best-studied ascidian enhancer is the a-element which is only 55bp. Furthermore, most ascidian enhancers have been found to be very close to the TSS, however in Drosophila, some have even been found 40kb from the TSS (Kvon et al., 2014), or even further (Ghavi-Helm et al., 2014). These differences in characteristics between the two taxa cannot be explained by genome compactness alone, as the two taxa are very similar in this respect.

ii. Studying endodermal GRNs more in depth

The work presented in this manuscript set out to study how *Ciona intestinalis* and *Phallusia mammaillata* have conserved their morphology during early embryogenesis despite extensive coding and non-coding sequence divergence. We decided to focus on the endodermal tissue from the 16-cell up to the 112-cell stage, during which time the endoderm precursor cells undergo two important evolutionarily conserved developmental processes: initial fate specification and early gastrulation.

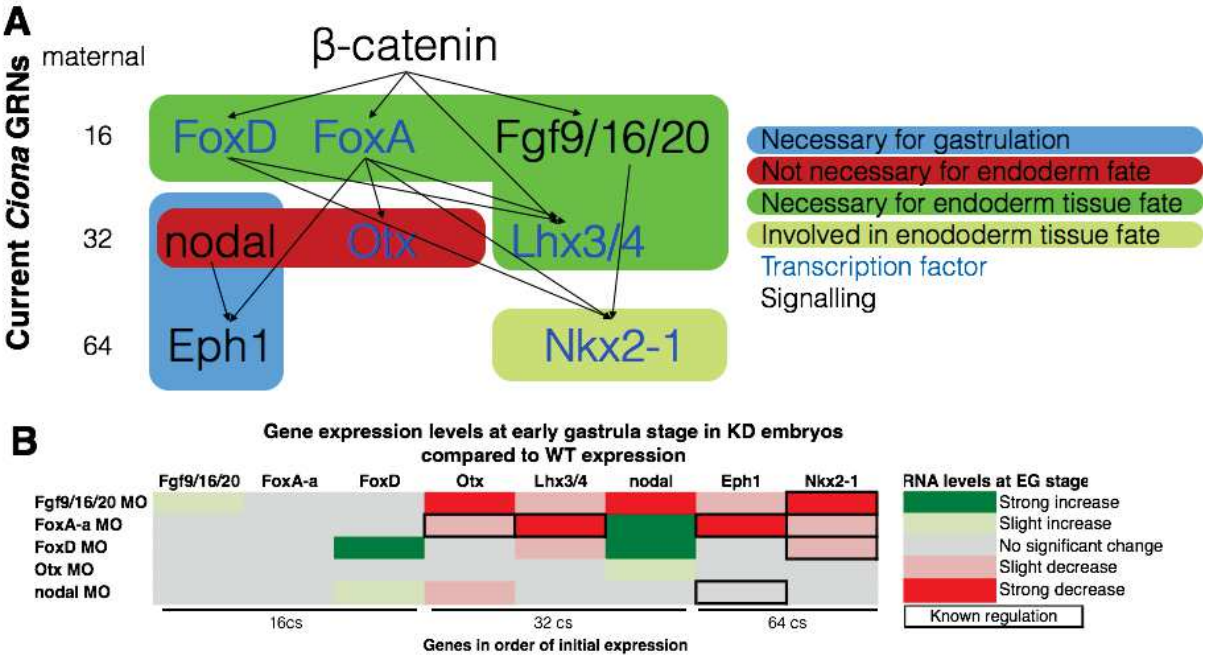


Figure 19. Ciona gene interactions. A) A schematic of what is currently known about the endodermal GRNs. B) A table showing increased or decreased gene expression levels in morpholino embryos compared to wild type expression levels at early gastrula stage, measured by Q-PCR (adapted from Imai et al., 2006).

Most of the work in *Ciona* that helped build the endodermal GRNs was based on extensive qualitative and quantitative expression data in morpholino experiments (Figure 19A). These experiments, however, do not suffice to determine direct interactions within the networks; for example, nodal knockdown seems to have no quantitative effect on Eph1 expression even though it is a direct regulator (Figure 19B).

We first compared by *in situ* hybridisation the transcriptional expression of orthologous regulatory genes in *Phallusia* and in *Ciona*. We then performed ATAC-seq in collaboration with the Jose Luis Gomez-Skarmeta lab in both *Ciona* and *Phallusia* to map chromatin accessibility through time in order to identify active enhancers. Once the endodermal enhancers were found, we aimed to identify TFBSs within the enhancers driving the expression of the genes in both species.

Bibliography

Abdul-Wajid, S., Veeman, M.T., Chiba, S., Turner, T.L., and Smith, W.C. (2014). Exploiting the extraordinary genetic polymorphism of ciona for developmental genetics with whole genome sequencing. Genetics *197*, 49–59.

Adelman, K., and Lis, J.T. (2012). Promoter-proximal pausing of RNA polymerase II: emerging roles in metazoans. Nat. Rev. Genet. *13*, 720–731.

Arensbergen, J. van, Steensel, B. van, and Bussemaker, H.J. (2014). In search of the determinants of enhancer–promoter interaction specificity. Trends Cell Biol. *24*, 695–702.

Arnold, C.D., Gerlach, D., Stelzer, C., Boryn, L.M., Rath, M., and Stark, A. (2013). Genome-wide quantitative enhancer activity maps identified by STARR-seq. Science *339*, 1074–1077.

Arnone, M.I., and Davidson, E.H. (1997). The hardwiring of development: organization and function of genomic regulatory systems. Dev. Camb. Engl. *124*, 1851–1864.

Arnosti, D.N., and Kulkarni, M.M. (2005). Transcriptional enhancers: Intelligent enhanceosomes or flexible billboards? J. Cell. Biochem. *94*, 890–898.

Artavanis-Tsakonas, S., Matsuno, K., and Fortini, M.E. (1995). Notch signaling. Science *268*, 225–232.

Aza-Blanc, P., Ramírez-Weber, F.A., Laget, M.P., Schwartz, C., and Kornberg, T.B. (1997). Proteolysis that is inhibited by hedgehog targets Cubitus interruptus protein to the nucleus and converts it to a repressor. Cell *89*, 1043–1053.

Barolo, S. (2012). Shadow enhancers: frequently asked questions about distributed cis-regulatory information and enhancer redundancy. BioEssays News Rev. Mol. Cell. Dev. Biol. *34*, 135–141.

Barthel, K.K.B., and Liu, X. (2008). A transcriptional enhancer from the coding region of ADAMTS5. PloS One *3*, e2184.

Beddington, R.S. (1994). Induction of a second neural axis by the mouse node. Dev. Camb. Engl. *120*, 613–620.

Bennett, D.C., and Lamoreux, M.L. (2003). The color loci of mice--a genetic century. Pigment Cell Res. *16*, 333–344.

Ben-Tabou de-Leon, S., and Davidson, E.H. (2007). Gene regulation: gene control network in development. Annu. Rev. Biophys. Biomol. Struct. *36*, 191.

Birney, E. (2012). An integrated encyclopedia of DNA elements in the human genome. Nature *489*, 57–74.

Blackwood, E.M., and Kadonaga, J.T. (1998). Going the distance: a current view of enhancer action. Science *281*, 60–63.

Britten, R.J., and Davidson, E.H. (1969). Gene regulation for higher cells: a theory. Science *165*, 349–357.
Bulger, M., and Groudine, M. (2011). Functional and mechanistic diversity of distal transcription enhancers. Cell *144*, 327–339.
Cadigan, K.M., and Nusse, R. (1997). Wnt signaling: a common theme in animal development. Genes Dev. *11*, 3286–3305.
Cande, J., Goltsev, Y., and Levine, M.S. (2009). Conservation of enhancer location in divergent insects. Proc. Natl. Acad. Sci. U. S. A. *106*, 14414–14419.
Cannavò, E., Khoueiry, P., Garfield, D.A., Geeleher, P., Zichner, T., Gustafson, E.H., Ciglar, L., Korbel, J.O., and Furlong, E.E.M. (2016). Shadow Enhancers Are Pervasive Features of Developmental Regulatory Networks. Curr. Biol. CB *26*, 38–51.
Carninci, P., Sandelin, A., Lenhard, B., Katayama, S., Shimokawa, K., Ponjavic, J., Semple, C.A.M., Taylor, M.S., Engström, P.G., Frith, M.C., et al. (2006). Genome-wide analysis of mammalian promoter architecture and evolution. Nat. Genet. *38*, 626–635.
Carroll, S.B. (2008). Evo-devo and an expanding evolutionary synthesis: a genetic theory of morphological evolution. Cell *134*, 25–36.
Carroll, S.B., Grenier, J.K., and Weatherbee, S.D. (2001). From DNA to Diversity: Molecular Genetics and the Evolution of Animal Design, 2nd Edition (Wiley-Blackwell).
Catarino, R.R., Neumayr, C., and Stark, A. (2017). Promoting transcription over long distances. Nat. Genet. *49*, 972–973.
Christiaen, L., Bourrat, F., and Joly, J.-S. (2005). A modular cis-regulatory system controls isoform-specific pitx expression in ascidian stomodaeum. Dev. Biol. *277*, 557–566.
Christiaen, L., Wagner, E., Shi, W., and Levine, M. (2009a). Microinjection of morpholino oligos and RNAs in sea squirt (Ciona) embryos. Cold Spring Harb. Protoc. *2009*, pdb.prot5347.
Christiaen, L., Wagner, E., Shi, W., and Levine, M. (2009b). Whole-mount in situ hybridization on sea squirt (Ciona intestinalis) embryos. Cold Spring Harb. Protoc. *2009*, pdb.prot5348.
Ciliberti, S., Martin, O.C., and Wagner, A. (2007). Innovation and robustness in complex regulatory gene networks. Proc. Natl. Acad. Sci. U. S. A. *104*, 13591–13596.
Comings, D.E. (1972). The structure and function of chromatin. Adv. Hum. Genet. *3*, 237–431.
Corbo, J.C., Levine, M., and Zeller, R.W. (1997). Characterization of a notochord-specific enhancer from the Brachyury promoter region of the ascidian, Ciona intestinalis. Dev. Camb. Engl. *124*, 589–602.
Crick, F. (1970). Central dogma of molecular biology. Nature *227*, 561–563.
Crocker, J., Abe, N., Rinaldi, L., McGregor, A.P., Frankel, N., Wang, S., Alsawadi, A., Valenti, P., Plaza, S., Payre, F., et al. (2015). Low affinity binding site clusters confer hox specificity and regulatory robustness. Cell *160*, 191–203.
Davidson, E.H., and Erwin, D.H. (2006). Gene regulatory networks and the evolution of animal body plans. Science *311*, 796–800.
De Robertis, E.M. (2006). Spemann's organizer and self-regulation in amphibian embryos. Nat. Rev. Mol. Cell Biol. *7*, 296–302.
Deaton, A.M., and Bird, A. (2011). CpG islands and the regulation of transcription. Genes Dev. *25*, 1010–1022.

Dehal, P., Satou, Y., Campbell, R.K., Chapman, J., Degnan, B., De Tomaso, A., Davidson, B., Di Gregorio, A., Gelpke, M., Goodstein, D.M., et al. (2002). The draft genome of Ciona intestinalis: insights into chordate and vertebrate origins. Science *298*, 2157–2167.
Delsuc, F., Brinkmann, H., Chourrout, D., and Philippe, H. (2006). Tunicates and not cephalochordates are the closest living relatives of vertebrates. Nature *439*, 965–968.
Denoeud, F., Henriet, S., Mungpakdee, S., Aury, J.-M., Da Silva, C., Brinkmann, H., Mikhaleva, J., Olsen, L.C., Jubin, C., Cañestro, C., et al. (2010). Plasticity of animal genome architecture unmasked by rapid evolution of a pelagic tunicate. Science *330*, 1381–1385.
Dhar, R., Weissman, S.M., Zain, B.S., Pan, J., and Lewis, A.M.J. (1974). The nucleotide sequence preceding an RNA polymerase initiation site on SV40 DNA. Part 2. The sequence of the early strand transcript. Nucleic Acids Res. *1*, 595–611.
Dixon, J.R., Selvaraj, S., Yue, F., Kim, A., Li, Y., Shen, Y., Hu, M., Liu, J.S., and Ren, B. (2012). Topological domains in mammalian genomes identified by analysis of chromatin interactions. Nature *485*, 376–380.
Duboule, D., and Dollé, P. (1989). The structural and functional organization of the murine HOX gene family resembles that of Drosophila homeotic genes. EMBO J. *8*, 1497–1505.
Dunham, I., Kundaje, A., Aldred, S.F., Collins, P.J., Davis, C.A., Doyle, F., Epstein, C.B., Frietze, S., Harrow, J., Kaul, R., et al. (2012). An integrated encyclopedia of DNA elements in the human genome. Nature *489*, 57–74.
Dunipace, L., Ozdemir, A., and Stathopoulos, A. (2011). Complex interactions between cis-regulatory modules in native conformation are critical for Drosophila snail expression. Dev. Camb. Engl. *138*, 4075–4084.
ENCODE Project Consortium (2012). An integrated encyclopedia of DNA elements in the human genome. Nature *489*, 57–74.
Erwin, D.H., and Davidson, E.H. (2009). The evolution of hierarchical gene regulatory networks. Nat. Rev. Genet. *10*, 141–148.
Farley, E.K., Olson, K.M., Zhang, W., Brandt, A.J., Rokhsar, D.S., and Levine, M.S. (2015). Suboptimization of developmental enhancers. Science *350*, 325–328.
Farley, E.K., Olson, K.M., Zhang, W., Rokhsar, D.S., and Levine, M.S. (2016). Syntax compensates for poor binding sites to encode tissue specificity of developmental enhancers. Proc. Natl. Acad. Sci. U. S. A. *113*, 6508–6513.
Femino, A.M., Fay, F.S., Fogarty, K., and Singer, R.H. (1998). Visualization of single RNA transcripts in situ. Science *280*, 585–590.
Frankel, N., Davis, G.K., Vargas, D., Wang, S., Payre, F., and Stern, D.L. (2010). Phenotypic robustness conferred by apparently redundant transcriptional enhancers. Nature *466*, 490–493.
Gannon, F., O'Hare, K., Perrin, F., LePennec, J.P., Benoist, C., Cochet, M., Breathnach, R., Royal, A., Garapin, A., Cami, B., et al. (1979). Organisation and sequences at the 5' end of a cloned complete ovalbumin gene. Nature *278*, 428–434.
Gaunt, S.J., and Paul, Y.-L. (2012). Changes in Cis-regulatory Elements during Morphological Evolution. Biology *1*, 557–574.
Gibcus, J.H., and Dekker, J. (2013). The hierarchy of the 3D genome. Mol. Cell *49*, 773–782.
Gisselbrecht, S.S., Barrera, L.A., Porsch, M., Aboukhalil, A., Estep, P.W., Vedenko, A., Palagi, A., Kim, Y., Zhu, X., Busser, B.W., et al. (2013). Highly parallel assays of tissue-specific enhancers in whole Drosophila embryos. Nat. Methods *10*, 774–780.

Gohl, D., Müller, M., Pirrotta, V., Affolter, M., and Schedl, P. (2008). Enhancer blocking and transvection at the Drosophila apterous locus. Genetics *178*, 127–143.
Gompel, N., and Prud'homme, B. (2009). The causes of repeated genetic evolution. Dev. Biol. *332*, 36–47.
Gompel, N., Prud'homme, B., Wittkopp, P.J., Kassner, V.A., and Carroll, S.B. (2005). Chance caught on the wing: cis-regulatory evolution and the origin of pigment patterns in Drosophila. Nature *433*, 481–487.
Graham, A., Papalopulu, N., and Krumlauf, R. (1989). The murine and Drosophila homeobox gene complexes have common features of organization and expression. Cell *57*, 367–378.
Graur, D. (2017). An Upper Limit on the Functional Fraction of the Human Genome. Genome Biol. Evol. *9*, 1880–1885.
Gray, S., and Levine, M. (1996). Transcriptional repression in development. Curr. Opin. Cell Biol. *8*, 358–364.
Gualdi, R., Bossard, P., Zheng, M., Hamada, Y., Coleman, J.R., and Zaret, K.S. (1996). Hepatic specification of the gut endoderm in vitro: cell signaling and transcriptional control. Genes Dev. *10*, 1670–1682.
Guenther, M.G., Levine, S.S., Boyer, L.A., Jaenisch, R., and Young, R.A. (2007). A Chromatin Landmark and Transcription Initiation at Most Promoters in Human Cells. Cell *130*, 77–88.
Haberle, V., and Lenhard, B. (2016). Promoter architectures and developmental gene regulation. Semin. Cell Dev. Biol. *57*, 11–23.
Hadzhiev, Y., Lang, M., Ertzer, R., Meyer, A., Strähle, U., and Müller, F. (2007). Functional diversification of sonic hedgehog paralog enhancers identified by phylogenomic reconstruction. Genome Biol. *8*, R106.
Haeussler, M., Jaszczyszyn, Y., Christiaen, L., and Joly, J.-S. (2010). A cis-regulatory signature for chordate anterior neuroectodermal genes. PLoS Genet. *6*, e1000912.
Halfon, M.S. (2017). Perspectives on Gene Regulatory Network Evolution. Trends Genet. TIG *33*, 436–447.
Harafuji, N., Keys, D.N., and Levine, M. (2002). Genome-wide identification of tissue-specific enhancers in the Ciona tadpole. Proc. Natl. Acad. Sci. U. S. A. *99*, 6802–6805.
Hare, E.E., Peterson, B.K., Iyer, V.N., Meier, R., and Eisen, M.B. (2008). Sepsid even-skipped enhancers are functionally conserved in Drosophila despite lack of sequence conservation. PLoS Genet. *4*, e1000106.
Harmston, N., Baresic, A., and Lenhard, B. (2013). The mystery of extreme non-coding conservation. Philos. Trans. R. Soc. Lond. B. Biol. Sci. *368*, 20130021.
He, B.Z., Holloway, A.K., Maerkl, S.J., and Kreitman, M. (2011). Does positive selection drive transcription factor binding site turnover? A test with Drosophila cis-regulatory modules. PLoS Genet. *7*, e1002053.
Heintzman, N.D., Stuart, R.K., Hon, G., Fu, Y., Ching, C.W., Hawkins, R.D., Barrera, L.O., Van Calcar, S., Qu, C., Ching, K.A., et al. (2007). Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. Nat. Genet. *39*, 311–318.
Hikosaka, A., Kusakabe, T., and Satoh, N. (1994). Short upstream sequences associated with the muscle-specific expression of an actin gene in ascidian embryos. Dev. Biol. *166*, 763–769.

Hill, M.M., Broman, K.W., Stupka, E., Smith, W.C., Jiang, D., and Sidow, A. (2008). The C. savignyi genetic map and its integration with the reference sequence facilitates insights into chordate genome evolution. Genome Res. *18*, 1369–1379.
Hirano, T., and Nishida, H. (2000). Developmental fates of larval tissues after metamorphosis in the ascidian, Halocynthia roretzi. II. Origin of endodermal tissues of the juvenile. Dev. Genes Evol. *210*, 55–63.
Hoekstra, H.E., and Nachman, M.W. (2003). Different genes underlie adaptive melanism in different populations of rock pocket mice. Mol. Ecol. *12*, 1185–1194.
Hong, J.-W., Hendrix, D.A., and Levine, M.S. (2008). Shadow enhancers as a source of evolutionary novelty. Science *321*, 1314.
Hoshino, Z., and Tokioka, T. (1967). An unusually robust Ciona from the northeastern coast of Honsyu island, Japan. Publ. SETO Mar. Biol. Lab. *15*, 275–290.
Hudson, C., Kawai, N., Negishi, T., and Yasuo, H. (2013). β-Catenin-driven binary fate specification segregates germ layers in ascidian embryos. Curr. Biol. CB *23*, 491–495.
Hudson, C., Sirour, C., and Yasuo, H. (2016). Co-expression of Foxa.a, Foxd and Fgf9/16/20 defines a transient mesendoderm regulatory state in ascidian embryos. eLife *5*.
Imai, K., Takada, N., Satoh, N., and Satou, Y. (2000). (beta)-catenin mediates the specification of endoderm cells in ascidian embryos. Dev. Camb. Engl. *127*, 3009–3020.
Imai, K.S., Hino, K., Yagi, K., Satoh, N., and Satou, Y. (2004). Gene expression profiles of transcription factors and signaling molecules in the ascidian embryo: towards a comprehensive understanding of gene networks. Dev. Camb. Engl. *131*, 4047–4058.
Imai, K.S., Levine, M., Satoh, N., and Satou, Y. (2006). Regulatory blueprint for a chordate embryo. Science *312*, 1183–1187.
Inukai, S., Kock, K.H., and Bulyk, M.L. (2017). Transcription factor-DNA binding: beyond binding site motifs. Curr. Opin. Genet. Dev. *43*, 110–119.
Iwafuchi-Doi, M., Donahue, G., Kakumanu, A., Watts, J.A., Mahony, S., Pugh, B.F., Lee, D., Kaestner, K.H., and Zaret, K.S. (2016). The Pioneer Transcription Factor FoxA Maintains an Accessible Nucleosome Configuration at Enhancers for Tissue-Specific Gene Activation. Mol. Cell *62*, 79–91.
Jack, J., Dorsett, D., Delotto, Y., and Liu, S. (1991). Expression of the cut locus in the Drosophila wing margin is required for cell type specification and is regulated by a distant enhancer. Dev. Camb. Engl. *113*, 735–747.
Jin, F., Li, Y., Dixon, J.R., Selvaraj, S., Ye, Z., Lee, A.Y., Yen, C.-A., Schmitt, A.D., Espinoza, C.A., and Ren, B. (2013). A high-resolution map of the three-dimensional chromatin interactome in human cells. Nature *503*, 290–294.
Junion, G., Spivakov, M., Girardot, C., Braun, M., Gustafson, E.H., Birney, E., and Furlong, E.E.M. (2012). A transcription factor collective defines cardiac cell fate and reflects lineage history. Cell *148*, 473–486.
Keys, D.N., Lee, B., Di Gregorio, A., Harafuji, N., Detter, J.C., Wang, M., Kahsai, O., Ahn, S., Zhang, C., Doyle, S.A., et al. (2005). A saturation screen for cis-acting regulatory DNA in the Hox genes of Ciona intestinalis. Proc. Natl. Acad. Sci. U. S. A. *102*, 679–683.
Kheradpour, P., Ernst, J., Melnikov, A., Rogov, P., Wang, L., Zhang, X., Alston, J., Mikkelsen, T.S., and Kellis, M. (2013). Systematic dissection of regulatory motifs in 2000 predicted human enhancers using a massively parallel reporter assay. Genome Res. *23*, 800–811.

Khoueiry, P., Rothbächer, U., Ohtsuka, Y., Daian, F., Frangulian, E., Roure, A., Dubchak, I., and Lemaire, P. (2010). A cis-regulatory signature in ascidians and flies, independent of transcription factor binding sites. Curr. Biol. CB *20*, 792–802.
Khoueiry, P., Girardot, C., Ciglar, L., Peng, P.-C., Gustafson, E.H., Sinha, S., and Furlong, E.E. (2017). Uncoupling evolutionary changes in DNA sequence, transcription factor occupancy and enhancer activity. eLife *6*.
Kikuta, H., Fredman, D., Rinkwitz, S., Lenhard, B., and Becker, T.S. (2007). Retroviral enhancer detection insertions in zebrafish combined with comparative genomics reveal genomic regulatory blocks - a fundamental feature of vertebrate genomes. Genome Biol. *8 Suppl 1*, S4.
Kimura, M. (1968). Genetic variability maintained in a finite population due to mutational production of neutral and nearly neutral isoalleles. Genet. Res. *11*, 247–269.
King, M.C., and Wilson, A.C. (1975). Evolution at two levels in humans and chimpanzees. Science *188*, 107–116.
Kintner, C.R., and Dodd, J. (1991). Hensen's node induces neural tissue in Xenopus ectoderm. Implications for the action of the organizer in neural induction. Dev. Camb. Engl. *113*, 1495–1505.
Kleinjan, D.A., and van Heyningen, V. (2005). Long-range control of gene expression: emerging mechanisms and disruption in disease. Am. J. Hum. Genet. *76*, 8–32.
de Kok, Y.J., Vossenaar, E.R., Cremers, C.W., Dahl, N., Laporte, J., Hu, L.J., Lacombe, D., Fischel-Ghodsian, N., Friedman, R.A., Parnes, L.S., et al. (1996). Identification of a hot spot for microdeletions in patients with X-linked deafness type 3 (DFN3) 900 kb proximal to the DFN3 gene POU3F4. Hum. Mol. Genet. *5*, 1229–1235.
Kowalevsky, A. (1886). Entwicklungsgeshichte der einfachen Ascidien. Mem. Acad. St Petersbourg *7*, 1–19.
Krivega, I., and Dean, A. (2012). Enhancer and promoter interactions—long distance calls. Curr. Opin. Genet. Dev. *22*, 79–85.
Kvon, E.Z., Kazmar, T., Stampfel, G., Yáñez-Cuna, J.O., Pagani, M., Schernhuber, K., Dickson, B.J., and Stark, A. (2014). Genome-scale functional characterization of Drosophila developmental enhancers in vivo. Nature *512*, 91–95.
Kwak, H., and Lis, J.T. (2013). Control of Transcriptional Elongation. Annu. Rev. Genet. *47*, 483–508.
Lagha, M., Bothma, J.P., and Levine, M. (2012). Mechanisms of transcriptional precision in animal development. Trends Genet. TIG *28*, 409–416.
Lai, F., and Shiekhattar, R. (2014). Enhancer RNAs: the new molecules of transcription. Curr. Opin. Genet. Dev. *25*, 38–42.
Lam, M.T.Y., Li, W., Rosenfeld, M.G., and Glass, C.K. (2014). Enhancer RNAs and regulated transcriptional programs. Trends Biochem. Sci. *39*, 170–182.
Lanctôt, C., Moreau, A., Chamberland, M., Tremblay, M.L., and Drouin, J. (1999). Hindlimb patterning and mandible development require the Ptx1 gene. Dev. Camb. Engl. *126*, 1805–1810.
Landry, C.R., Lemos, B., Rifkin, S.A., Dickinson, W.J., and Hartl, D.L. (2007). Genetic properties influencing the evolvability of gene expression. Science *317*, 118–121.
Lang, G., Gombert, W.M., and Gould, H.J. (2005). A transcriptional regulatory element in the coding sequence of the human Bcl-2 gene. Immunology *114*, 25–36.

Le Petillon, Y., Luxardi, G., Scerbo, P., Cibois, M., Leon, A., Subirana, L., Irimia, M., Kodjabachian, L., Escriva, H., and Bertrand, S. (2017). Nodal/Activin Pathway is a Conserved Neural Induction Signal in Chordates. Nat. Ecol. Evol. *1*, 1192–1200.
Ledford, H. CRISPR studies muddy results of older gene research. Nat. News.
Lemaire, P. (2011). Evolutionary crossroads in developmental biology: the tunicates. Dev. Camb. Engl. *138*, 2143–2152.
Lettice, L.A., Horikoshi, T., Heaney, S.J.H., Baren, M.J. van, Linde, H.C. van der, Breedveld, G.J., Joosse, M., Akarsu, N., Oostra, B.A., Endo, N., et al. (2002). Disruption of a long-range cis-acting regulator for Shh causes preaxial polydactyly. Proc. Natl. Acad. Sci. *99*, 7548–7553.
Lettice, L.A., Heaney, S.J.H., Purdie, L.A., Li, L., de Beer, P., Oostra, B.A., Goode, D., Elgar, G., Hill, R.E., and de Graaff, E. (2003). A long-range Shh enhancer regulates expression in the developing limb and fin and is associated with preaxial polydactyly. Hum. Mol. Genet. *12*, 1725–1735.
Levo, M., and Segal, E. (2014). In pursuit of design principles of regulatory sequences. Nat. Rev. Genet. *15*, 453–468.
Ludwig, M.Z., Manu, null, Kittler, R., White, K.P., and Kreitman, M. (2011). Consequences of eukaryotic enhancer architecture for gene expression dynamics, development, and fitness. PLoS Genet. *7*, e1002364.
Lynch, V.J., and Wagner, G.P. (2008). Resurrecting the role of transcription factor change in developmental evolution. Evol. Int. J. Org. Evol. *62*, 2131–2154.
Marcil, A., Dumontier, E., Chamberland, M., Camper, S.A., and Drouin, J. (2003). Pitx1 and Pitx2 are required for development of hindlimb buds. Dev. Camb. Engl. *130*, 45–55.
Markstein, M., Markstein, P., Markstein, V., and Levine, M. (2002). Genome-wide analysis of clustered Dorsal binding sites identifies putative target genes in the Drosophila embryo. Proc Natl Acad Sci USA *99*, 763–768.
Mathelier, A., Fornes, O., Arenillas, D.J., Chen, C.-Y., Denay, G., Lee, J., Shi, W., Shyr, C., Tan, G., Worsley-Hunt, R., et al. (2016). JASPAR 2016: a major expansion and update of the open-access database of transcription factor binding profiles. Nucleic Acids Res. *44*, D110-115.
Maurano, M.T., Humbert, R., Rynes, E., Thurman, R.E., Haugen, E., Wang, H., Reynolds, A.P., Sandstrom, R., Qu, H., Brody, J., et al. (2012). Systematic localization of common disease-associated variation in regulatory DNA. Science *337*, 1190–1195.
McKeown, A.N., Bridgham, J.T., Anderson, D.W., Murphy, M.N., Ortlund, E.A., and Thornton, J.W. (2014). Evolution of DNA specificity in a transcription factor family produced a new gene regulatory module. Cell *159*, 58–68.
Miele, A., and Dekker, J. (2008). Long-range chromosomal interactions and gene regulation. Mol. Biosyst. *4*, 1046–1057.
Mullen, A.C., Orlando, D.A., Newman, J.J., Lovén, J., Kumar, R.M., Bilodeau, S., Reddy, J., Guenther, M.G., DeKoter, R.P., and Young, R.A. (2011). Master transcription factors determine cell-type-specific responses to TGF-β signaling. Cell *147*, 565–576.
Mundy, N.I. (2005). A window on the genetics of evolution: MC1R and plumage colouration in birds. Proc. Biol. Sci. *272*, 1633–1640.
Nam, J., and Davidson, E.H. (2012). Barcoded DNA-tag reporters for multiplex cis-regulatory analysis. PloS One *7*, e35934.
Nishida, H. (1987). Cell lineage analysis in ascidian embryos by intracellular injection of a tracer enzyme. III. Up to the tissue restricted stage. Dev. Biol. *121*, 526–541.

Nishida, H. (1996). Vegetal egg cytoplasm promotes gastrulation and is responsible for specification of vegetal blastomeres in embryos of the ascidian Halocynthia roretzi. Dev. Camb. Engl. *122*, 1271–1279.
Nitta, K.R., Jolma, A., Yin, Y., Morgunova, E., Kivioja, T., Akhtar, J., Hens, K., Toivonen, J., Deplancke, B., Furlong, E.E.M., et al. (2015). Conservation of transcription factor binding specificities across 600 million years of bilateria evolution. eLife *4*.
Oda-Ishii, I., Kubo, A., Kari, W., Suzuki, N., Rothbächer, U., and Satou, Y. (2016). A Maternal System Initiating the Zygotic Developmental Program through Combinatorial Repression in the Ascidian Embryo. PLoS Genet. *12*, e1006045.
Oetting, W.S., Garrett, S.S., Brott, M., and King, R.A. (2005). P gene mutations associated with oculocutaneous albinism type II (OCA2). Hum. Mutat. *25*, 323.
Ohler, U., and Wassarman, D.A. (2010). Promoting developmental transcription. Dev. Camb. Engl. *137*, 15–26.
Olson-Manning, C.F., Wagner, M.R., and Mitchell-Olds, T. (2012). Adaptive evolution: evaluating empirical support for theoretical predictions. Nat. Rev. Genet. *13*, 867–877.
Orenstein, Y., and Shamir, R. (2014). A comparative analysis of transcription factor binding models learned from PBM, HT-SELEX and ChIP data. Nucleic Acids Res. *42*, e63.
Otto, S.P. (2004). Two steps forward, one step back: the pleiotropic effects of favoured alleles. Proc. Biol. Sci. *271*, 705–714.
Paixão, T., and Azevedo, R.B.R. (2010). Redundancy and the evolution of cis-regulatory element multiplicity. PLoS Comput. Biol. *6*, e1000848.
Panne, D. (2008). The enhanceosome. Curr. Opin. Struct. Biol. *18*, 236–242.
Panne, D., Maniatis, T., and Harrison, S.C. (2007). An Atomic Model of the Interferon-β Enhanceosome. Cell *129*, 1111–1123.
Papatsenko, D., and Levine, M. (2007). A rationale for the enhanceosome and other evolutionarily constrained enhancers. Curr. Biol. CB *17*, R955-957.
Parker, H.G., VonHoldt, B.M., Quignon, P., Margulies, E.H., Shao, S., Mosher, D.S., Spady, T.C., Elkahloun, A., Cargill, M., Jones, P.G., et al. (2009). An expressed fgf4 retrogene is associated with breed-defining chondrodysplasia in domestic dogs. Science *325*, 995–998.
Pavlicev, M., and Wagner, G.P. (2012). A model of developmental evolution: selection, pleiotropy and compensation. Trends Ecol. Evol. *27*, 316–322.
Pennati, R., Ficetola, G.F., Brunetti, R., Caicci, F., Gasparini, F., Griggio, F., Sato, A., Stach, T., Kaul-Strehlow, S., Gissi, C., et al. (2015). Morphological Differences between Larvae of the Ciona intestinalis Species Complex: Hints for a Valid Taxonomic Definition of Distinct Species. PloS One *10*, e0122879.
Pérez-Lluch, S., Blanco, E., Tilgner, H., Curado, J., Ruiz-Romero, M., Corominas, M., and Guigó, R. (2015). Absence of canonical marks of active chromatin in developmentally regulated genes. Nat. Genet. *47*, 1158–1167.
Perry, M.W., Cande, J.D., Boettiger, A.N., and Levine, M. (2009). Evolution of insect dorsoventral patterning mechanisms. Cold Spring Harb. Symp. Quant. Biol. *74*, 275–279.
Peter, I.S., and Davidson, E.H. (2011a). A gene regulatory network controlling the embryonic specification of endoderm. Nature *474*, 635–639.
Peter, I.S., and Davidson, E.H. (2011b). Evolution of gene regulatory networks controlling body plan development. Cell *144*, 970–985.
Phillips-Cremins, J.E. (2014). Unraveling architecture of the pluripotent genome. Curr. Opin. Cell Biol. *28*, 96–104.

Pribnow, D. (1975). Nucleotide sequence of an RNA polymerase binding site at an early T7 promoter. Proc. Natl. Acad. Sci. U. S. A. *72*, 784–788.
Protas, M.E., Hersey, C., Kochanek, D., Zhou, Y., Wilkens, H., Jeffery, W.R., Zon, L.I., Borowsky, R., and Tabin, C.J. (2006). Genetic analysis of cavefish reveals molecular convergence in the evolution of albinism. Nat. Genet. *38*, 107–111.
Rebeiz, M., Reeves, N.L., and Posakony, J.W. (///). SCORE: A computational approach to the identification of cis-regulatory modules and target genes in whole-genome sequence data. Proc Natl Acad Sci USA.
Rebeiz, M., Pool, J.E., Kassner, V.A., Aquadro, C.F., and Carroll, S.B. (2009). Stepwise modification of a modular enhancer underlies adaptation in a Drosophila population. Science *326*, 1663–1667.
Ristoratore, F., Spagnuolo, A., Aniello, F., Branno, M., Fabbrini, F., and Di Lauro, R. (1999). Expression and functional analysis of Cititf1, an ascidian NK-2 class gene, suggest its role in endoderm development. Dev. Camb. Engl. *126*, 5149–5159.
Roure, A., Lemaire, P., and Darras, S. (2014). An otx/nodal regulatory signature for posterior neural development in ascidians. PLoS Genet. *10*, e1004548.
Sardet, C., Paix, A., Prodon, F., Dru, P., and Chenevert, J. (2007). From oocyte to 16-cell stage: cytoplasmic and cortical reorganizations that pattern the ascidian embryo. Dev. Dyn. Off. Publ. Am. Assoc. Anat. *236*, 1716–1731.
Satoh, N. (2014). Developmental Genomics of Ascidians (Wiley-Blackwell).
Satou, Y., Imai, K.S., and Satoh, N. (2001). Early embryonic expression of a LIM-homeobox gene Cs-lhx3 is downstream of beta-catenin and responsible for the endoderm differentiation in Ciona savignyi embryos. Dev. Camb. Engl. *128*, 3559–3570.
Satou, Y., Kawashima, T., Shoguchi, E., Nakayama, A., and Satoh, N. (2005). An integrated database of the ascidian, Ciona intestinalis: towards functional genomics. Zoolog. Sci. *22*, 837–843.
Sayou, C., Monniaux, M., Nanao, M.H., Moyroud, E., Brockington, S.F., Thévenon, E., Chahtane, H., Warthmann, N., Melkonian, M., Zhang, Y., et al. (2014). A promiscuous intermediate underlies the evolution of LEAFY DNA binding specificity. Science *343*, 645–648.
Schaffner, W. (2015). Enhancers, enhancers - from their discovery to today's universe of transcription enhancers. Biol. Chem. *396*, 311–327.
Schier, A.F., and Talbot, W.S. (2005). Molecular genetics of axis formation in zebrafish. Annu. Rev. Genet. *39*, 561–613.
Segal, E., Raveh-Sadka, T., Schroeder, M., Unnerstall, U., and Gaul, U. (2008). Predicting expression patterns from regulatory sequence in Drosophila segmentation. Nature *451*, 535–540.
Shapiro, M.D., Marks, M.E., Peichel, C.L., Blackman, B.K., Nereng, K.S., Jónsson, B., Schluter, D., and Kingsley, D.M. (2004). Genetic and developmental basis of evolutionary pelvic reduction in threespine sticklebacks. Nature *428*, 717–723.
Shapiro, M.D., Bell, M.A., and Kingsley, D.M. (2006). Parallel genetic origins of pelvic reduction in vertebrates. Proc. Natl. Acad. Sci. U. S. A. *103*, 13753–13758.
Shen, Y., Yue, F., McCleary, D.F., Ye, Z., Edsall, L., Kuan, S., Wagner, U., Dixon, J., Lee, L., Lobanenkov, V.V., et al. (2012). A map of the cis-regulatory sequences in the mouse genome. Nature *488*, 116–120.

Sherrard, K., Robin, F., Lemaire, P., and Munro, E. (2010). Sequential activation of apical and basolateral contractility drives ascidian endoderm invagination. Curr. Biol. CB *20*, 1499–1510.
Shlyueva, D., Stampfel, G., and Stark, A. (2014). Transcriptional enhancers: from properties to genome-wide predictions. Nat. Rev. Genet. *15*, 272–286.
Spitz, F., Gonzalez, F., and Duboule, D. (2003). A global control region defines a chromosomal regulatory landscape containing the HoxD cluster. Cell *113*, 405–417.
Stampfel, G., Kazmar, T., Frank, O., Wienerroither, S., Reiter, F., and Stark, A. (2015). Transcriptional regulators form diverse groups with context-dependent regulatory functions. Nature *528*, 147–151.
Steiner, C.C., Römpler, H., Boettger, L.M., Schöneberg, T., and Hoekstra, H.E. (2009). The genetic basis of phenotypic convergence in beach mice: similar pigment patterns but different genes. Mol. Biol. Evol. *26*, 35–45.
Stern, D.L., and Orgogozo, V. (2008). The loci of evolution: how predictable is genetic evolution? Evol. Int. J. Org. Evol. *62*, 2155–2177.
Stern, D.L., and Orgogozo, V. (2009). Is genetic evolution predictable? Science *323*, 746–751.
Stolfi, A., Gandhi, S., Salek, F., and Christiaen, L. (2014). Tissue-specific genome editing in Ciona embryos by CRISPR/Cas9. Dev. Camb. Engl. *141*, 4115–4120.
Struhl, K., and Segal, E. (2013). Determinants of nucleosome positioning. Nat. Struct. Mol. Biol. *20*, 267–273.
Suryamohan, K., Hanson, C., Andrews, E., Sinha, S., Scheel, M.D., and Halfon, M.S. (2016). Redeployment of a conserved gene regulatory network during Aedes aegypti development. Dev. Biol. *416*, 402–413.
Sutter, N.B., Bustamante, C.D., Chase, K., Gray, M.M., Zhao, K., Zhu, L., Padhukasahasram, B., Karlins, E., Davis, S., Jones, P.G., et al. (2007). A single IGF1 allele is a major determinant of small size in dogs. Science *316*, 112–115.
Szeto, D.P., Rodriguez-Esteban, C., Ryan, A.K., O'Connell, S.M., Liu, F., Kioussi, C., Gleiberman, A.S., Izpisúa-Belmonte, J.C., and Rosenfeld, M.G. (1999). Role of the Bicoid-related homeodomain factor Pitx1 in specifying hindlimb morphogenesis and pituitary development. Genes Dev. *13*, 484–494.
Takatori, N., Kumano, G., Saiga, H., and Nishida, H. (2010). Segregation of germ layer fates by nuclear migration-dependent localization of Not mRNA. Dev. Cell *19*, 589–598.
Tokuhiro, S.-I., Tokuoka, M., Kobayashi, K., Kubo, A., Oda-Ishii, I., and Satou, Y. (2017). Differential gene expression along the animal-vegetal axis in the ascidian embryo is maintained by a dual functional protein Foxd. PLoS Genet. *13*, e1006741.
Treen, N., Yoshida, K., Sakuma, T., Sasaki, H., Kawai, N., Yamamoto, T., and Sasakura, Y. (2014). Tissue-specific and ubiquitous gene knockouts by TALEN electroporation provide new approaches to investigating gene function in Ciona. Dev. Camb. Engl. *141*, 481–487.
Trompouki, E., Bowman, T.V., Lawton, L.N., Fan, Z.P., Wu, D.-C., DiBiase, A., Martin, C.S., Cech, J.N., Sessa, A.K., Leblanc, J.L., et al. (2011). Lineage regulators direct BMP and Wnt pathways to cell-specific programs during differentiation and regeneration. Cell *147*, 577–589.
True, J.R., and Haag, E.S. (2001). Developmental system drift and flexibility in evolutionary trajectories. Evol. Dev. *3*, 109–119.

Villar, D., Berthelot, C., Aldridge, S., Rayner, T.F., Lukk, M., Pignatelli, M., Park, T.J., Deaville, R., Erichsen, J.T., Jasinska, A.J., et al. (2015). Enhancer evolution across 20 mammalian species. Cell *160*, 554–566.
Wagner, A. (2005). Robustness, evolvability, and neutrality. FEBS Lett. *579*, 1772–1778.
Wagner, A. (2012). The role of robustness in phenotypic adaptation and innovation. Proc. Biol. Sci. *279*, 1249–1258.
Wingender, E., Schoeps, T., and Dönitz, J. (2013). TFClass: an expandable hierarchical classification of human transcription factors. Nucleic Acids Res. *41*, D165-170.
Wittkopp, P.J., True, J.R., and Carroll, S.B. (2002). Reciprocal functions of the Drosophila yellow and ebony proteins in the development and evolution of pigment patterns. Dev. Camb. Engl. *129*, 1849–1858.
Yang, C., Bolotin, E., Jiang, T., Sladek, F.M., and Martinez, E. (2007). Prevalence of the Initiator over the TATA box in human and yeast genes and identification of DNA motifs enriched in human TATA-less core promoters. Gene *389*, 52–65.
Young, N.M., and Hallgrímsson, B. (2005). Serial homology and the evolution of mammalian limb covariation structure. Evol. Int. J. Org. Evol. *59*, 2691–2704.
Young, N.M., Wagner, G.P., and Hallgrímsson, B. (2010). Development and the evolvability of human limbs. Proc. Natl. Acad. Sci. U. S. A. *107*, 3400–3405.
Young, N.M., Winslow, B., Takkellapati, S., and Kavanagh, K. (2015). Shared rules of development predict patterns of evolution in vertebrate segmentation. Nat. Commun. *6*, ncomms7690.
Yuan, G.-C., Cai, L., Elowitz, M., Enver, T., Fan, G., Guo, G., Irizarry, R., Kharchenko, P., Kim, J., Orkin, S., et al. (2017). Challenges and emerging directions in single-cell analysis. Genome Biol. *18*, 84.
Zaret, K.S., and Carroll, J.S. (2011). Pioneer transcription factors: establishing competence for gene expression. Genes Dev. *25*, 2227–2241.
Zhang, Y., Wong, C.-H., Birnbaum, R.Y., Li, G., Favaro, R., Ngan, C.Y., Lim, J., Tai, E., Poh, H.M., Wong, E., et al. (2013). Chromatin connectivity maps reveal dynamic promoter–enhancer long-range associations. Nature *504*, 306–310.

Evolution of the cis-regulatory architecture between embryos of the divergent *Phallusia mammillata* and *Ciona intestinalis* ascidians

Alicia Madgwick[1], Marta Nagri[2], Christelle Dantec[1], Damien Gailly[1], Ulla-Maj-Fiuza[1], Jose Luis Gomez-Skarmeta[2], and Patrick Lemaire[1]

1) Centre de Recherche en Biologie cellulaire de Montpellier, Unversité de Montpellier, CNRS, 1919 route de Mende, F-34293, Montpellier cedex 5, France.

2) Centro Andaluz de Biología del Desarrollo (CABD), Consejo Superior de Investigaciones Científicas/Universidad Pablo de Olavide, Carretera de Utrera Km1, Sevilla, Spain

Abstract

How can embryonic morphogenesis be evolutionarily conserved in spite of extensive divergence in coding and non-coding genome sequences? To address this question, we worked on the early development of two ascidians who diverged several hundred million years ago, *Phallusia mammillata* and *Ciona intestinalis*. These species share almost identical early morphogenesis and stereotyped cell lineages. Remarkably, however, their genomes are divergent to the extent that their non-coding sequences cannot be aligned and gene order has not been conserved. We focused our attention on the evolution of the gene regulatory networks driving the fate specification and invagination of pre-gastrula endodermal precursors. We first show by *in situ* hybridisation that the endodermal expression of these genes is qualitatively conserved between *Phallusia* and in *Ciona*.

To study how these genes conserved their regulation in spite of extensive non-coding sequence divergence, we mapped using ATAC-seq the chromatin accessibility landscapes in both species to identify active regulatory regions. 32 of the 36 chromatin-accessible regions for our set of endodermal genes behaved as active regulatory sequences when tested by embryo electroporation, 17 of these regions acted during the pregastrula stages. These sequences include examples of pleiotropic enhancers and shadow enhancers, suggesting that in spite of the compactness of its genome and its peculiar stereotyped mode of development, *Phallusia* uses similar regulatory principles as other metazoans. 17 of the tested sequences had conserved *cis*-regulatory activity in both species in spite of sequence divergence. We have identified putative transcription factor binding sites in endodermal enhancers in both species to identify conserved upstream regulators shared between *Phallusia* and *Ciona*.

Taken together our results suggest that extensive transcription factor binding site turn over, without pervasive change in GRNs architecture, may explain the qualitative conservation of gene expression patterns between highly divergent ascidian genomes.

I. Introduction

Development and morphogenesis are precisely orchestrated by complex Gene Regulatory Networks. These networks combine transcription factors, and the *cis*-regulatory sequences to which they bind to control gene expression. Morphological change during evolution is frequently caused by variations in the transcriptional programme, and in particular in cis-regulatory sequences (1), The converse is, however, not always true and there are many examples of divergent gene regulatory networks producing very similar phenotypic outputs, a scenario referred to as developmental systems drift or DSD (2). Examples of DSD include vulva specification in nematodes (3) and heart morphogenesis in ascidians (4). Our understanding of the complex relationships between genotype and phenotype currently remains too fragmentary to predict the phenotypic outcome of a regulatory mutation.

Ascidians are marine invertebrate chordates belonging to the vertebrate sister group, the tunicates (5). Like vertebrates, their embryos develop through a tadpole larval stage. Unlike vertebrates, however, ascidian embryonic development is highly stereotyped, and proceeds with invariant cell lineages. Each cell can be individually named and found across all embryos of a given species and the number, names and location of its progeny is also precisely defined. Strikingly, even distantly related species, which diverged more than 300 million years ago share very similar or identical cell lineages. This provides a rigorous framework to compare, with cellular resolution, the developmental programmes across species. Paradoxically, ascidian genomes are highly divergent (4, 6, 7). Both coding and non-coding sequences evolve rapidly within the group, gene orders are frequently scrambled along chromosomes, and even highly conserved gene clusters, such as the Hox cluster, are exploded (8). Ascidians thus constitute a very interesting taxon to study DSD.

Cis-regulatory sequences are DNA segments, which act as binding platforms for transcription factors, and control the expression of genes. These sequences are so divergent between distant ascidian species that homologous sequences can generally not be aligned anymore (4, 9, 10). In some cases, highly divergent regulatory sequences have retained the ability to respond to the same transcription factor combination (9, 10), the accumulation of cis-regulatory mutations thus leaving the GRN architecture intact. In other cases, however, a functional *cis*-regulatory sequence from one species can become "unintelligible" to another species (4, 11). The conservation of the function of *cis*-regulatory sequences has been assayed in too few cases to estimate the relative frequency of conservation vs divergence of *cis*-regulatory activity. The relative contributions of local TF binding site turn over versus more global GRN rewiring between ascidian species remains unknown.

In this study, we focused our attention on eight key nodes of the early *Ciona* mesendodermal GRN, driving two important evolutionary-conserved developmental processes: initial mesendoderm fate specification and the onset of endoderm invagination during gastrulation. We first systematically compared the expression of these eight genes between two distant phlebobranch species, *Phallusia mammillata* and *Ciona intestinalis*. We then used ATAC-seq (12) to identify accessible chromatin regions in the vicinity of these genes in both species and show that these sequences include the main *cis*-regulatory sequences driving their endodermal expression. We finally compared the activity of these sequences between species. These analyses brought to light a surprising level of conservation of chromatin features and cis-regulatory logic between these species despite extreme sequence divergence.

II. Results

1. Regulatory states in *Ciona* and *Phallusia* through time

We first investigated the level of conservation of orthologous regulatory gene expression patterns in early *Ciona* and *Phallusia* endodermal precursors up to the 112-cells stage by which point the endoderm tissue fate has been determined and gastrulation is initiated. *Ciona* expression profiles of most transcription factors and signalling molecules have been extensively characterised by *in situ* hybridization (13, 14). We mined the ANISEED database (6), that catalogues expression data from published work, to define a set of regulatory genes expressed zygotically throughout all endoderm precursor cells for at least one of the following stages: 16-, 32-, 64- or 112-cell stages, and therefore likely to have a pan-endodermal role. This set of 15 genes includes 9 transcription factors genes and 6 genes involved in signalling, whose *Phallusia mammillata* orthologs were retrieved in ANISEED, and individually validated. 14 *Ciona* genes had an unambiguous 1-to-1 ortholog in *Phallusia mammillata* (*FoxAa*, *Nodal*, *Nkx2-1*, *Lhx3/4*, *Otx*, *Tolloid*, *SoxB1*, *Lefty*, *FGF9/16/20*, *Hes-b*, *ADMP*, *Eph1*, *Zf(C3H)*, *Elk*). *FoxD* has undergone a very recent duplication in *Ciona robusta*. The

paralogs are highly similar (96% identity at the nucleotide level) and difficult to distinguish by molecular approaches. They correspond to a unique *P. mammillata* gene.

Next, we determined the pregastrula spatial and temporal expression patterns of the *Phallusia* regulatory genes by *in situ* hybridization and compared them to their individual *Ciona* orthologs at the equivalent stages (Supp. Figure S1A), as illustrated for *FoxAa* in Figure 1A. All 15 *Ciona* and *Phallusia* orthologous pairs were found to share expression within endodermal precursor cells at one or more stages in development (Supp. Figure S1B). Conservationn of expression was overall high, yet varied between perfect cell-to-cell conservation of the expression pattern to significant divergence in up to 50% of cells (Figure 1B). Cases of divergence may reflect evolutionary change, but could also reflect either slight differences in the stages studied in *Ciona* and *Phallusia* or slight experimental variation, the *Ciona* and *Phallusia* patterns originating from distinct laboratories.

The regulatory state of a territory is defined as the combination of regulatory genes expressed in the territory. To explore the conservation of the endodermal regulatory states at successive stages between *Ciona* and *Phallusia*, we designed a stage-by-stage regulatory state similarity score (see methods) and applied it to each cell. Figure 1C illustrates that pan-endodermal regulatory states are overall very well conserved, and that the genes studied are sufficient to define an unambiguous endodermal identity.

This qualitative description of expression patterns was complemented by the semi-quantitative comparison of the level and dynamics of expression of these genes in staged RNA-seq experiments carried out on whole embryos of both species (Brozovic et al., submitted) (Supp. Fig. S2). Expression dynamics was overall conserved for Lhx3, FoxA, Ttf1, SoxB1, and Zf(c3h), while other genes showed heterochronies of activation between the two species, which were unlikely to be due to experimental error as the exact same stages were collected. Normalized expression values could also differ between the two species.

Overall, we conclude that early endodermal regulatory states are qualitatively well conserved between *Ciona* and *Phallusia*, suggesting the conservation of ancestral phlebobranchian endodermal regulatory states in spite of extensive genomic sequence divergence. The precise kinetics and level of expression of individual regulatory genes may, however, vary between species.

2. Chromatin features at regulatory regions

Two main scenarios have been put forward to explain the evolutionary conservation of regulatory states. The first scenario proposes a global conservation of GRN architecture, defined by the regulatory links established between transcription factors and the *cis*-regulatory sequences they bind to. This scenario predicts that the combination of transcription factors binding to each *cis*-regulatory sequence operating within the network should be conserved. By contrast the second scenario proposes that extensive

neutral evolution of these GRN is also compatible with regulatory state conservation (15). In this scenario, one expects that combinations of transcription factors binding to orthologous cis-regulatory sequences may be specific to each species. To discriminate between these two models, we set out to identify and compare the activity of the *cis*-regulatory sequences driving the expression of our set of *Phallusia* endodermal genes.

An open chromatin state is a prerequisite for the accessibility of most cis-regulatory regions to transcription factors (Ramachandran and Henikoff 2016). We first established ATAC-seq (12) to map open chromatin regions on whole *Ciona* embryos at the 64- and 112-cell stages. As expected, most of the open chromatin regions were located in regions of high sequence conservation between *Ciona robusta* and *Ciona savignyi* (Fig. 2A). Previously characterized *Ciona cis*-regulatory sequences active before the onset of gastrulation overlapped with open chromatin regions, as exemplified by the analysis of the *Ciona Otx, brachyury and snail* loci (Figure 2A and Supp. Figure S3). We next compared the chromatin accessibility landscapes in embryos in which most cells have been converted to endoderm (not shown) by pharmacological inhibition of GSK3 with the CHIR99021 inhibitor from the 8-cell stage. Although the effect was moderate, GSK3 inhibition increased the accessibility of known endoderm enhancers, and decreased that of ectodermal enhancers (Supp. Fig. S4). Finally, we did not detect major differences between the 64- and 112-cell chromatin landscapes. Thus, is *Ciona* open chromatin regions detected by ATAC-seq are enriched in *cis*-regulatory regions and chromatin accessibility is insufficient to predict the spatio-temporal pattern of expression of neighboring genes.

These initial results in *Ciona* encouraged us to build an ATAC-seq resource for *Phallusia mammillata*. We performed ATAC-seq in *Phallusia mammillata* whole wild-type embryos at the 16-, 32-, 64- and 112-cell stages. The sequence of most open chromatin regions was highly conserved between *Phallusia mammillata* and *Phallusia fumigata* (Figure 2B). As in *Ciona*, treatment of embryos with the pharmacological GSK3-inhibitor CHIR99021 from the 8-cell stage (4µM) only had a modest impact on chromatin accessibility landscapes (Figure 2B, compare 64-cell WT and GSK3 inhib; Supp. Figure S5). Finally, chromatin became accessible ahead of the onset of expression of the genes studied and were overall very similar at the 4 stages analysed, even in the vicinity of genes showing dynamic zygotic expression (e.g. TTF1, Figure S5). Using MACS2, we extracted 38500 statistically significant peaks of accessible chromatin in WT 64-cell embryos. Peaks had an average length of 180bp, compatible with the estimated size of ascidian enhancers. They covered a total of 6.9 Mb of coding and non coding sequecnes or 3% of the *Phallusia* genome draft. Peaks were on all scaffolds, with an average distance between peaks of 1.7kb (Supp. Fig. S6).

3. Enhancer screen in *Phallusia*

We next used the *Phallusia mammillata* ATAC-seq data to search for the regulatory sequences driving the endodermal expression of our set of 10 *Phallusia mammillata* endodermal regulatory genes.

We considered that enhancer boundaries corresponded to the boundaries of the peaks of accessible chromatin and scanned up to 6kb of non-coding sequences upstream of the gene transcription start site, plus the intronic regions. This identified 24 single-peak candidate *cis*-regulatory regions ranging from 220bp to 600bp, plus 12 multi-peak regions. Each of these regions was placed upstream of the stable NLS-LacZ reporter gene, with or without a minimal promoter, and electroporated into either *Phallusia* or *Ciona* fertilized eggs (16). We first assayed reporter gene activity at the larval stage: because of the stability of the LacZ protein and the rapidity of ascidian development, LacZ activity detected at the larval stage summarises the activity of the construct during all embryonic stages (Figure 2C).

Of the 36 non-coding open chromatin regions that were tested for regulatory activity at the larval stage, 32 were found to have spatially-restricted *cis*-regulatory activity, including 23 distal enhancers and 8 proximal elements combining enhancer and promoter activity (Figure 2D). This high success rate suggests that most peaks of open chromatin detected during ascidian embryogenesis encode autonomous *cis*-regulatory activity that can be detected out of its endogenous context in reporter assays. On average, there is at least 1 enhancer per kb in the non-coding regions upstream of the genes tested.

To identify the early endodermal *cis*-regulatory regions active prior to the onset of gastrulation, we reanalyzed at the gastrula stage the activity of sequences scoring positive at the larval stage (Figure 2D and Supp. Figure S5). This identified 17 early *cis*-regulatory sequences, 8 of which were active in the endoderm. All sequences with early onset of activity were in proximal position, within under 1kb of the TSS. Interestingly, the chromatin of the loci that were inactive before the 112-cell stage was open many stages ahead of their activation, strengthening the notion that chromatin accessibility is not sufficient for *cis*-regulatory activity.

4. Pleiotropic enhancers and shadow enhancers

Enhancer pleiotropy, the reuse of cis-regulatory loci at different times and in different tissues, has been observed in *Drosophila* and has been proposed to play an important role in the evolution of regulatory landscapes during the emergence of evolutionary novelties (17, 18). One of our enhancers, the FGF9/16/20-Pm05 enhancer was active at the 112-cell stage in A-line posterior neural territories. By the larval stage, however, LacZ activity is primarily detected in the endoderm, notochord and muscle (Figure 3A). Ascidians thus also appear to make use of pleiotropic enhancers and the prevalence of such activities will be interesting to assess in the future.

Studies in *Drosophila* and vertebrates suggest that the expression of regulatory genes is often driven by apparently redundant (or shadow) enhancers (19), which have been proposed to increase the robustness of the developmental program to genetic and environmental variations (20, 21). Ascidians are among the most polymorphic metazoan (22) and live in environments strongly affected by variations in temperature or salinity. Several shadow enhancers have previously been identified in *Ciona* (9, 23). We found 3

pairs of partially redundant enhancers driving the early expression in *Eph1*, *Otx* and *Fgf9/16/20* (Figures 3B, C); the fact that at least 6 of our 17 early enhancers act redundantly at early stages suggests that shadow enhancers may be a frequent feature of the *cis*-regulatory logic of ascidian regulatory genes.

These 3 pairs of shadow enhancers share few structural characteristics. The Eph1 enhancers are separated by two exons of Eph1, as one of the enhancers is located intronically; the Otx enhancers are separated by another early enhancer and the third pair of shadow enhancers, driving Fgf9/16/20 expression, are next to one another within just 1220bp. The Eph1 shadow enhancers both drive expression in the endoderm; Eph1Pm02 additionally driving expression in the notochord whereas Eph1Pm03 is active in B7.5 and B7.7 that will give rise to the mesenchyme, the TVCs and muscle. The Otx shadow enhancers have overlapping activity in B7.5, B7.7, B8.15 and B8.7 that will mostly produce muscle; OtxPm05 additionally drives expression in the B-line endoderm, B7.1 and B7.2, along with A8.7 and A8.8 that will give rise to the nervous system. The Pfgf9/16/20 shadow enhancers both drive expression in A8.7, A8.8, A8.15 and A8.16 that will give rise to the nervous system and muscle, however, only Fgf9/16/20 has additional activity in B8.7 and B8.8 muscle cells.

5. Conservation of enhancer location and activity between *Ciona* and *Phallusia*.

Finally, we tested the conservation of the *cis*-regulatory logic between *Phallusia* and *Ciona*. We first compared the chromatin accessibility landscapes around orthologous genes between the two species. Orthologous genes often seem to share very similar open chromatin regions at the same distance from the gene (Figure 4). For example, there are three accessible regions just upstream of the Otx gene in *Ciona*, contained within the sequences A, B and C, (Figure 4A) that look very similar to the three regions upstream of Otx in Phallusia, within equivalent regions A, B and C (Figure 4B). We next wanted to compare the activity of these *Ciona* sequences to the equivalent *Phallusia* enhancers.

The three early *Otx* enhancers with the same open chromatin patterns seem to have a very similar activity as in *Ciona* driving early expression in several vegetal cells (Figure 4). Therefore, we tested the orthologous region in *Ciona* to those found in *Phallusia* upstream of Lhx3/4. The Lhx3/4 enhancer from both species drives expression in the majority of the vegetal cells, including the endoderm, reproducing the 32-cell stage expression of Lhx3/4. The equivalent regions in the two species were found to have the same qualitative activity and therefore behave as orthologous enhancers.

Orthologous *Ciona* and *Phallusia* enhancer sequences between cannot be aligned. To investigate whether the cis-regulatory logic of *Phallusia* sequences could be understood by *Ciona* and vice versa despite sequence divergence we tested 17 *Phallusia* regulatory sequences in *Ciona*. 14 of these 17 sequences were found to have the same activity in both species (Figure S6). The regulatory sequences found to drive early endoderm expression were also found to have conserved cis-activity such as in *Lhx3/4* (Figure 5A).

We wanted to find the upstream regulators of the enhancers driving early endodermal expression. Seeing as the *trans*- and *cis*-logic seems to have been conserved, we looked for the same binding sites within the orthologous Lhx3/4 *Ciona* and *Phallusia* enhancers. Due to the fact that Phallusia and Ciona are unalignable, we aligned them to more closely related species, *Phallusia mammillata* to *Phallusia fumigata* and *Ciona robusta* to *Ciona savygnyi*. We then searched for putative binding sites that are conserved within the genera and whose position has mostly been conserved also. We decided to focus on the TFBSs for genes known that regulate many early vegetal genes: Tcf, Ets and Fox (Figure 5B). The orthologous Lhx3/4 enhancers share binding sites for these TFs, however, with significant TFBS turnover.

III. Discussion

Ascidians have notoriously fast evolving genomes; *Phallusia* and *Ciona* have very divergent genomes to the point that only their coding regions are alignable. Despite their remarkable resemblance throughout early embryogenesis, Ciona and Phallusia have considerably different expression levels in orthologous genes. Previous work found that they have equally different gene expression profiles between 50% of orthologous genes across time (personal communication J. Piette). We wanted to see to what extent the gene expression of regulatory genes was qualitatively divergent at the cellular level. Similarly to other ascidian work, we found that their expression patterns at the cellular level were very similar, however, we did find slight differences. The gene expression levels in whole embryo did not necessarily reflect these differences.

Therefore, we wanted to know if the GRNs driving their expression would also be different seeing as their non-coding regions are unalignable. We decided to focus on the GRNs driving the endodermal fate differentiation and gastrulation. We investigated the regulatory sequences driving the expression of these genes and found that they share some remarkable characteristics. As anticipated, the enhancers are more conserved than other non-coding regions. A particularly striking feature was that many of the late enhancers were actually open many hours before the onset of their activity. Chromatin accessibility is not sufficient to determine that an enhancer is currently activity because many of the enhancers in this study were open much earlier than their onset of activity. Lineage determining transcription factor availability may be the limiting factors, the chromatin landscape acting as a "canalizer" for the binding of these factors.

Several enhancer characteristics are shared not only across ascidians but also across species. The earliest enhancers in *Phallusia* and *Ciona* were found closest to the promoter; this had been seen in Drosophila (24).

We found a high number of enhancers around the genes located quite densely upstream of the gene they are activating. This number is even more surprising enhancers can have a pleiotropic activity, used several times throughout early development in different tissues. Although we found an enhancer with pleiotropic activity, this high number of

enhancers could be due to shadow enhancers. Shadow enhancers, also known as redundant enhancers, are rather prevalent in Drosophila and can play different role in regulating expression (25). They were more frequent than expected and than what was previously seen in ascidians. Very few shadow enhancers are known in ascidians (Matsumoto 2008?, Irvine 2011?, Farley 2016). We found a third of the ealy enhancers are shadow enhancers. Shadow enhancers have been associated with robustness of expression, more commonly in genes involved in development.

The regulatory landscapes across ascidians seem rather well conserved. This demonstrates microsynteny because there has been a loss of gene synteny different between these species. A more marked different was found within the enhancers themselves which seem to have undergone extensive/considerable TF shuffling. This suggests that changes within ascidian sequences are due to micro- rather than macro-evolution. Further work on ascidian shadow enhancers would be interesting to see if they promote robustness for spatial and temporal expression only and if expression levels fluctuate because this is not under selective pressure.

IV. Material and methods

Animal origin and embryo manipulation
Phallusia mammillata and *Ciona intestinalis* (previously *Ciona intestinalis* type B) were provided by the Centre de Ressources Biologiques Marines in Roscoff . *Ciona intestinalis* dechorionation was done in either fertilised or unfertilised eggs as previously described and *Phallusia mammillata* unfertilised eggs were dechorionated in 0,4% trypsin in ASW with gentle shaking for 2 hours. All live experiments were performed with these dechorionated embryos. Embryos were, where indicated, treated with 4 µM of CHIR99021 from the 8-cell stage.

Gene models (names, unique gene IDs and genome assembly)
The genes analysed in this project were from the *Ciona robusta* KH2012 assembly, represented by the following gene models: *Foxa.a* KH.C11.313, *Foxd* KH.C8.396, *nodal* KH.L106.16, *Nkx2-1* KH.C10.338, *Lhx3/4* KH.S215.4, *Otx* KH.C4.84, *Eph1* KH.C1.404, *Tolloid* KH.C12.156, *Lefty1/2* KH.C3.411, *Fgf9/16/*20 KH.C2.125, SoxB1 KH.C1.99, *Hes-b* KH.C3.312, *ADMP* KH.C2.421, Zf(C3H) KH.C4.182, *Elk* KH.C8.247. For *Phallusia mammillata*, the gene models were as follows: *Foxa.a* Phmamm.CG.MTP2014.S50.g01891, *Foxd* Phmamm.CG.MTP2014.S253.g06179, *nodal* Phmamm.CG.MTP2014.S1711.g15500, *Nkx2-1* Phmamm.CG.MTP2014.S597.g10419, *Lhx3/4* Phmamm.CG.MTP2014.S2332.g16546, *Otx* Phmamm.CG.S166.MTP2014.g04580, *Eph1* Phmamm.CG.MTP2014.S160.g04451, *Tolloid* Phmamm.CG.MTP2014.S14.g00600, *Lefty* Phmamm.CG.MTP2014.S362.g07801, *Fgf9/16/*20 Phmamm.CG.MTP2014.S128.g03805, *SoxB1* Phmamm.CG.MTP2014.S272.g06489, *Hes-b* Phmamm.CG.MTP2014.S180.g04861, *ADMP* Phmamm.CG.MTP2014.S74.g02578, *Zf(C3H)* Phmamm.CG.MTP2014.S224.g05660, *Elk* Phmamm.CG.MTP2014.S411.g08422

In situ hybridisation
Wholemount *in situ* hybridisation was performed as previously described for both *Ciona robusta* and *Phallusia mammillata* (26). Dig-labelled probes for *Phallusia mammillata* were synthesised from the following cDNAs: *Otx* (AHC0AAA196YH07 or AHC0AAA223YP12), *Lefty* (AHC0AAA31YI04), *Nhx2-1* (AHC0AAA32YD05 or AHC0AAA267YK08), *Tolloid* (AHC0AAA14YA03), *Fgf9/16/20* (AHC0AAA33YA03 or AHC0AAA22YD05), *Nodal* (AHC0AAA149YE05). The *Eph1* probe was generated by PCR from cDNA (a kind gift from Dr. UM Fiuza).

ATAC-seq assays, sequence preprocessing, alignment and peak calling
ATAC-seq assays were carried out essentially as described (12) starting with 50-70 thousand embryonic cells. The embryos were not dissociated prior to nucleus extraction. Sequencing was done an an Illumina HiSeq2500 (BGI) and around 20 million sequence reads obtained per sample. Illumina adapters were trimmed by cutadapt, alignment was done with bowtie2, (very-sensitive parameter). Only a minority of sequences aligned to the nuclear genome (*Ciona* ~30%; *Phallusia* ~17%). significant Peaks were extracted with MACS2 with the following parameters: gsize=187138625 (genome size = genome mappable size excluding repeated sequences) --nomodel --shift -94 --extsize 188. SEquences were submitted to the Short Read Archive (project number XXXX)

Gateway and electroporation of the regulatory sequences

Constructs to test the activity of the regulatory sequences from both species were generated by Gateway cloning into a destination vector including the FoxAa promoter from *Ciona robusta,* which drives expression throughout the vegetal cells from the 16-cell stage. The primers for the FoxA promoter and each of the non-coding sequences can be found in SUPP. The genomic DNA used to amplify by PCR the genomic fragments for *Ciona robusta* was a kind gift from Dr. Nietta Spagnuolo (Naples). *Phallusia mammillata* genomic DNA was extracted from Roscoff animals.

Electroporation was done as previously described (27), with a single pulse of 16ms at 50V for *Ciona intestinalis* and two pulses of 10ms at 40V for *Phallusia mammillata*. X-gal staining was performed as previously described (27).

in silico TFBS predictions
Putative transcription factor binding sites were chosen within enhancers that were aligned to a closely related species and if the binding site was well conserved. *Phallusia mammillata* regulatory sequences were aligned to *Phallusia fumigata* and *Ciona robusta* sequences to *Ciona savignyi* using BLASTn. The conserved putative binding sites within *Phallusia* or *Ciona* genera were searched for using ConTra v1 (28). Only binding sites found in all four species were considered for further testing. The sequences with mutated sites were ordered from GeneArt (ThermoFisher).

ACKNOWLEDGEMENT:

BIBLIOGRAPHY

1. Stern,D.L. and Orgogozo,V. (2008) The loci of evolution: how predictable is genetic evolution? *Evol. Int. J. Org. Evol.*, **62**, 2155–2177.
2. True,J.R. and Haag,E.S. (2001) Developmental system drift and flexibility in evolutionary trajectories. *Evol. Dev.*, **3**, 109–119.
3. Sommer,R.J. (2012) Evolution of regulatory networks: nematode vulva induction as an example of developmental systems drift. *Adv. Exp. Med. Biol.*, **751**, 79–91.
4. Stolfi,A., Lowe,E.K., Racioppi,C., Ristoratore,F., Brown,C.T., Swalla,B.J. and Christiaen,L. (2014) Divergent mechanisms regulate conserved cardiopharyngeal development and gene expression in distantly related ascidians. *eLife*, **3**, e03728.
5. Lemaire,P. (2011) Evolutionary crossroads in developmental biology: the tunicates. *Dev. Camb. Engl.*, **138**, 2143–2152.
6. Brozovic,M., Martin,C., Dantec,C., Dauga,D., Mendez,M., Simion,P., Percher,M., Laporte,B., Scornavacca,C., Di Gregorio,A., *et al.* (2016) ANISEED 2015: a digital framework for the comparative developmental biology of ascidians. *Nucleic Acids Res.*, **44**, D808-818.
7. Tsagkogeorga,G., Cahais,V. and Galtier,N. (2012) The population genomics of a fast evolver: high levels of diversity, functional constraint and molecular adaptation in the tunicate Ciona intestinalis. *Genome Biol. Evol.*, 10.1093/gbe/evs054.
8. Duboule,D. (2007) The rise and fall of Hox gene clusters. *Dev. Camb. Engl.*, **134**, 2549–2560.
9. Oda-Ishii,I., Bertrand,V., Matsuo,I., Lemaire,P. and Saiga,H. (2005) Making very similar embryos with divergent genomes: conservation of regulatory mechanisms of Otx between the ascidians Halocynthia roretzi and Ciona intestinalis. *Dev. Camb. Engl.*, **132**, 1663–74.
10. Roure,A., Lemaire,P. and Darras,S. (2014) An otx/nodal regulatory signature for posterior neural development in ascidians. *PLoS Genet.*, **10**, e1004548.
11. Takahashi,H., Mitani,Y., Satoh,G. and Satoh,N. (1999) Evolutionary alterations of the minimal promoter for notochord-specific Brachyury expression in ascidian embryos. *Dev. Camb. Engl.*, **126**, 3725–3734.
12. Buenrostro,J.D., Giresi,P.G., Zaba,L.C., Chang,H.Y. and Greenleaf,W.J. (2013) Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nat. Methods*, **10**, 1213–1218.
13. Imai,K.S., Hino,K., Yagi,K., Satoh,N. and Satou,Y. (2004) Gene expression profiles of transcription factors and signaling molecules in the ascidian embryo: towards a comprehensive understanding of gene networks. *Dev. Camb. Engl.*, **131**, 4047–4058.
14. Imai,K.S., Levine,M., Satoh,N. and Satou,Y. (2006) Regulatory blueprint for a chordate embryo. *Science*, **312**, 1183–1187.
15. Ciliberti,S., Martin,O.C. and Wagner,A. (2007) Innovation and robustness in complex regulatory gene networks. *Proc. Natl. Acad. Sci. U. S. A.*, **104**, 13591–13596.

16. Corbo,J.C., Levine,M. and Zeller,R.W. (1997) Characterization of a notochord-specific enhancer from the Brachyury promoter region of the ascidian, Ciona intestinalis. *Dev. Camb. Engl.*, **124**, 589–602.
17. Rebeiz,M., Jikomes,N., Kassner,V.A. and Carroll,S.B. (2011) Evolutionary origin of a novel gene expression pattern through co-option of the latent activities of existing regulatory sequences. *Proc. Natl. Acad. Sci. U. S. A.*, **108**, 10036–10043.
18. Rebeiz,M. and Tsiantis,M. (2017) Enhancer evolution and the origins of morphological novelty. *Curr. Opin. Genet. Dev.*, **45**, 115–123.
19. Cannavò,E., Khoueiry,P., Garfield,D.A., Geeleher,P., Zichner,T., Gustafson,E.H., Ciglar,L., Korbel,J.O. and Furlong,E.E.M. (2016) Shadow Enhancers Are Pervasive Features of Developmental Regulatory Networks. *Curr. Biol. CB*, **26**, 38–51.
20. Perry,M.W., Boettiger,A.N., Bothma,J.P. and Levine,M. (2010) Shadow enhancers foster robustness of Drosophila gastrulation. *Curr. Biol. CB*, **20**, 1562–1567.
21. Frankel,N., Erezyilmaz,D.F., McGregor,A.P., Wang,S., Payre,F. and Stern,D.L. (2011) Morphological evolution caused by many subtle-effect substitutions in regulatory DNA. *Nature*, **474**, 598–603.
22. Nydam,M.L. and Harrison,R.G. (2010) Polymorphism and divergence within the ascidian genus Ciona. *Mol. Phylogenet. Evol.*, **56**, 718–726.
23. Farley,E.K., Olson,K.M., Zhang,W., Rokhsar,D.S. and Levine,M.S. (2016) Syntax compensates for poor binding sites to encode tissue specificity of developmental enhancers. *Proc. Natl. Acad. Sci. U. S. A.*, **113**, 6508–6513.
24. Blythe,S.A. and Wieschaus,E.F. (2016) Establishment and maintenance of heritable chromatin structure during early Drosophila embryogenesis. *eLife*, **5**.
25. Cannavò,E., Khoueiry,P., Garfield,D.A., Geeleher,P., Zichner,T., Gustafson,E.H., Ciglar,L., Korbel,J.O. and Furlong,E.E.M. (2016) Shadow Enhancers Are Pervasive Features of Developmental Regulatory Networks. *Curr. Biol. CB*, **26**, 38–51.
26. Christiaen,L., Wagner,E., Shi,W. and Levine,M. (2009) Whole-mount in situ hybridization on sea squirt (Ciona intestinalis) embryos. *Cold Spring Harb. Protoc.*, **2009**, pdb.prot5348.
27. Bertrand,V., Hudson,C., Caillol,D., Popovici,C. and Lemaire,P. (2003) Neural tissue in ascidian embryos is induced by FGF9/16/20, acting via a combination of maternal GATA and Ets transcription factors. *Cell*, **115**, 615–27.
28. Hooghe,B., Hulpiau,P., van Roy,F. and De Bleser,P. (2008) ConTra: a promoter alignment analysis tool for identification of transcription factor binding sites across species. *Nucleic Acids Res.*, **36**, W128-132.

FIGURES

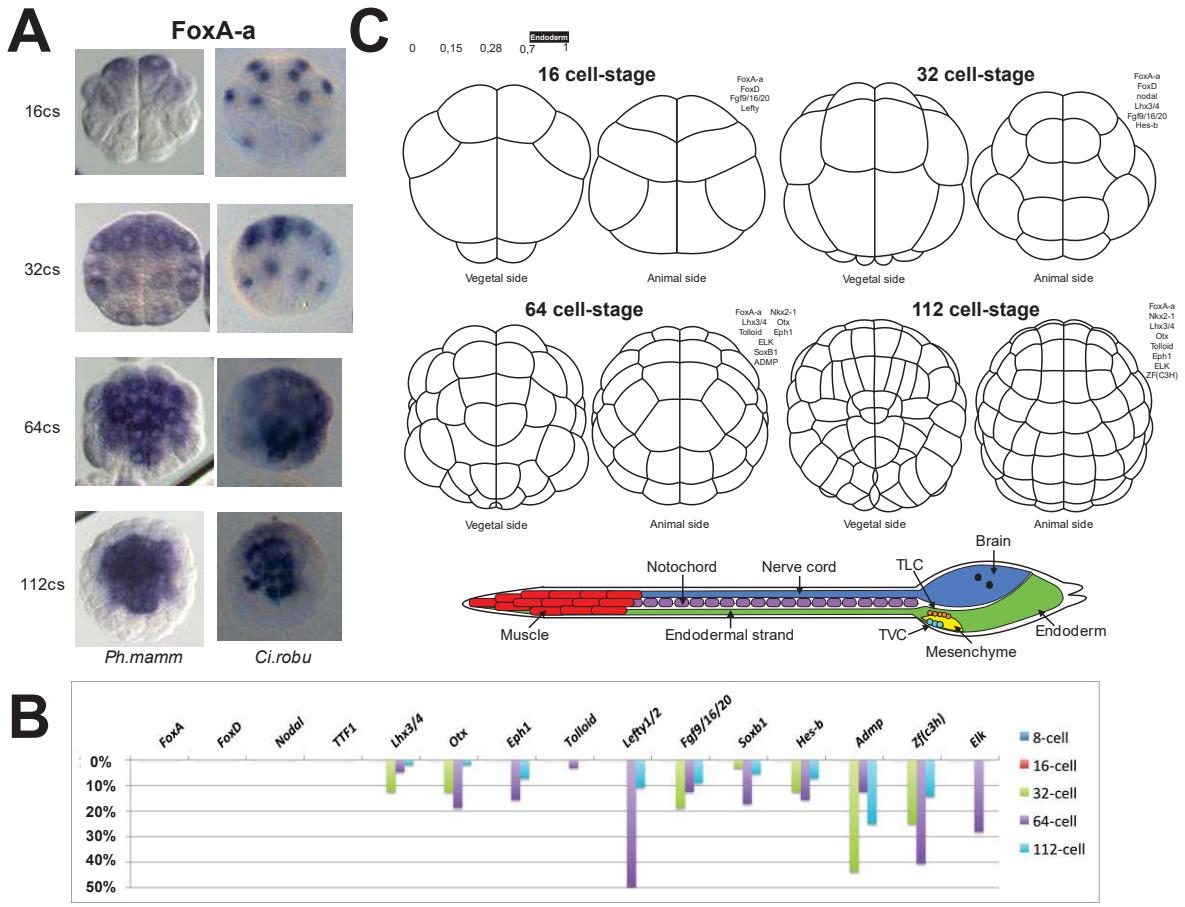


Figure 1: Conservation of regulatory states between *Ciona intestinalis* and *Phallusia mammillata.* A) Example of perfectly conserved expression patterns at the 16-, 32-, 64- and 112-cell stage of *Ciona* and *Phallusia* orthologs: the FoxAa situation. B) Quantification of the conservation of expression patterns of the studied genes in each developmental stage. The values indicated by colors the fraction of cells at each stage for which the expression status (on/off) differs between the two species. C) Conservation of the endodermal regulatory states between *Ciona* and *Phallusia*. left half of each embryo schema: The colors indicate for each cell at each stage the score of conservation between regulatory states defined by the indicated genes: (2x(# genes with conserved expression) + (# genes expressed in a single species))/ 2x(# total genes). This score reflects both the ancestral origin of extant regulatory states, and the divergence of these regulatory states in non endodermal territories. Tissue fates are indicated on the right half of each embryo schema.

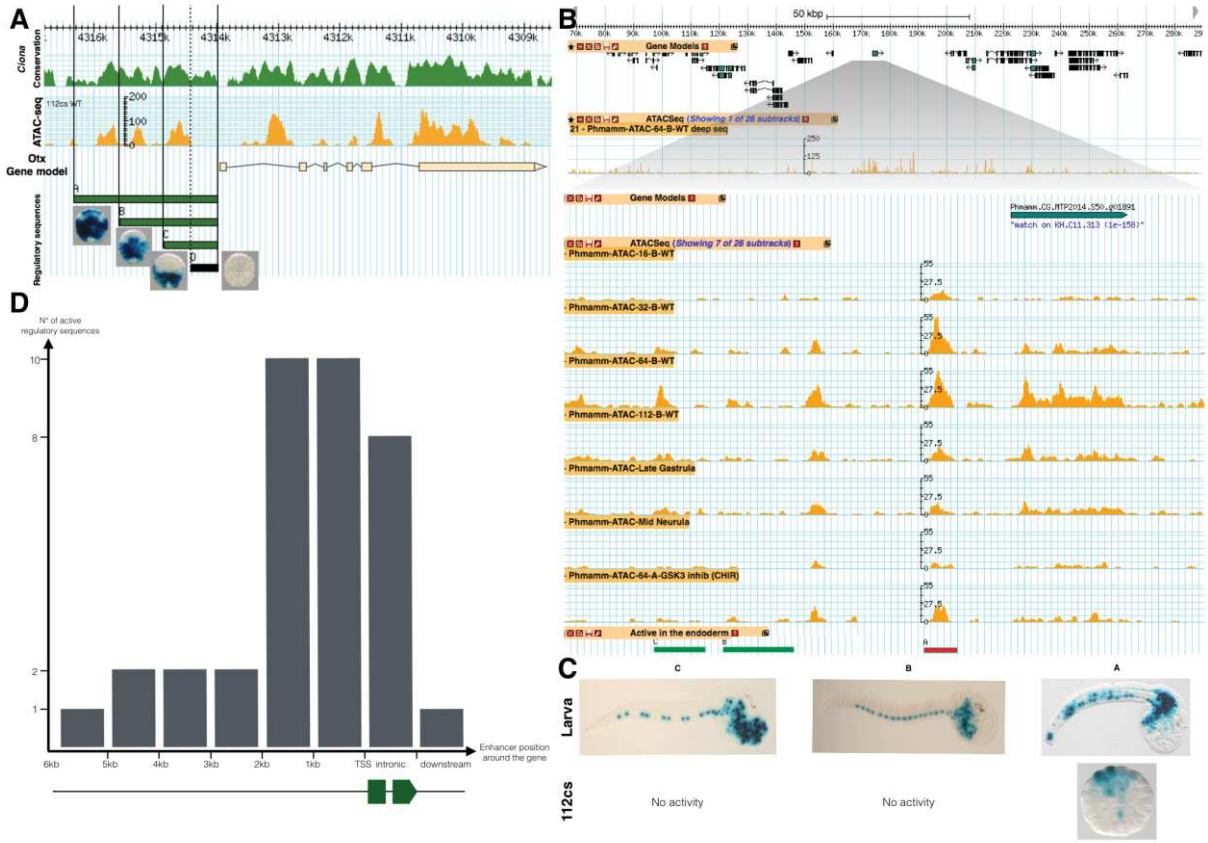


Figure 2: ATAC-seq efficiently identifies cis-regulatory sequences in Ciona and Phallusia. A) *Ciona* chromatin accessibility landscape at the *Otx* locus. From top to bottom. First track (green): local sequence conservation between *Ciona robusta* and *Ciona savignyi*. Second track (Orange): ATAC-seq pattern at the 112-cell stage in whole wildtype embryos. Third track: position of the *Otx* exons. Bottom track: location of the genomic constructs tested by electroporation, with their pattern of activity (adapted from Bertrand et al, 2003). B) *Phallusia* chromatin accessibility map at consecutive developmental stages around the FoxAa gene. Top: 220kb view centered on the FoxAa locus revealing a higher density of ATAC-seq peak in the vicinity of *FoxAa*. Bottom panel: zoomed view of the FoxAa locus. From top to bottom: first track, FoxAa single-exon model; second track(orange) , ATAC-seq landscape in whole wild-type embryos at the 16-, 32-, 64-, 112-cell, late gastrula and late neurula stages, and in CHIR-treated embryos at the 64-cell stages; third track, position of the three tested candidate regulatory regions. ATAC-seq values are not precisely normalized between stages. C) patterns of reporter gene activity at the hatching larval and 112-cell stages following electroporation of the indicated construct. D) Position of active constructs with respect to the TSS. Y-axis: # of active sequences. X-axis: distance to TSS and gene body.

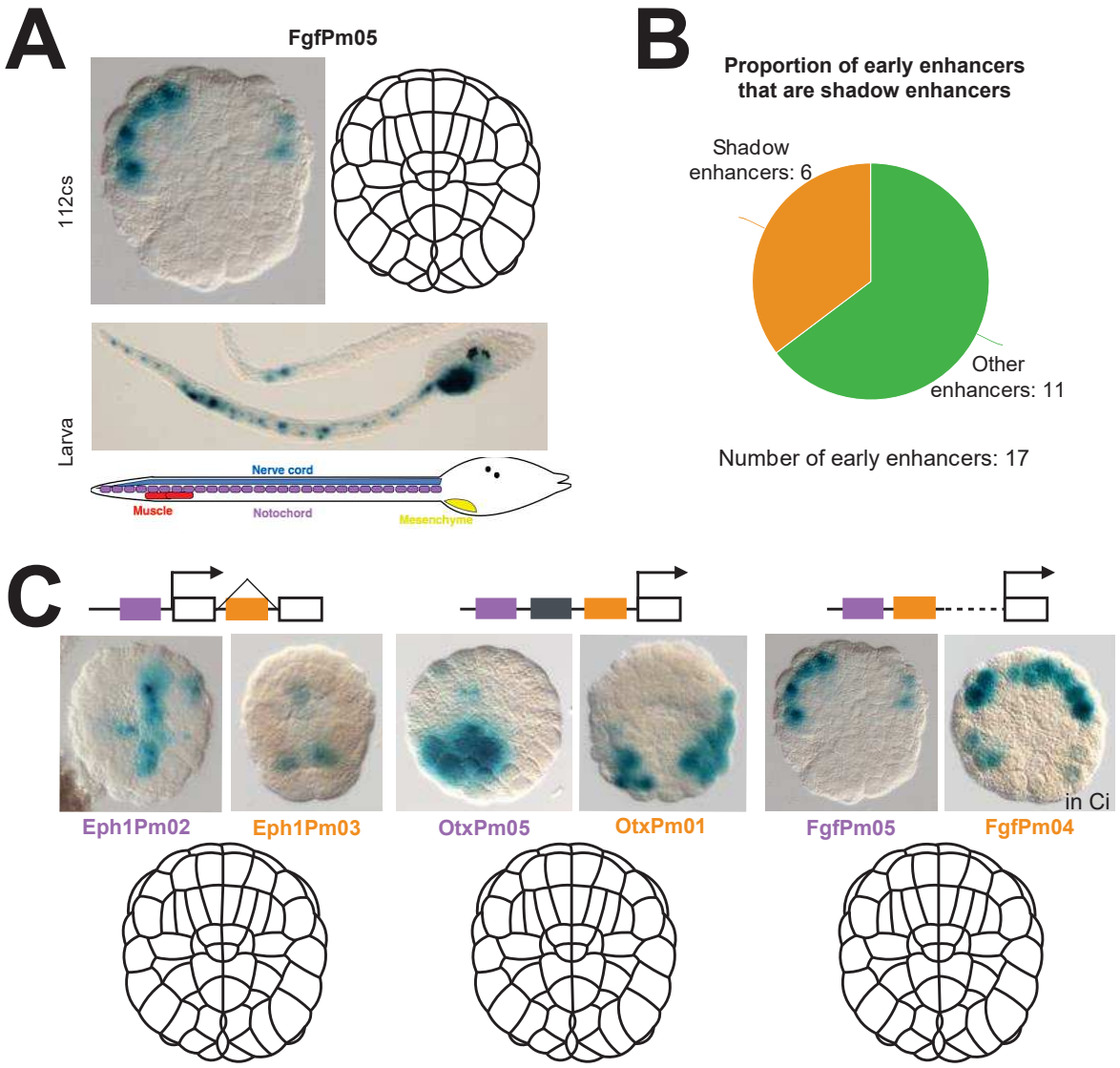


Figure 3: Pleiotropic and shadow enhancers. A) Pleiotropy of the FGF9/16/20 pm5 enhancer (see Figure S5). Top panels: detected activity in the nerve cord lineages at the 112-cell stage. Bottom panels: detected activity in the nerve cord, notochord and muscle lineages at the larval stages. B) Fraction of shadow enhancers identified among the early (gastrula) enhancers. C) The 6 shadow enhancers identified, with their expression patterns at the early gastrula stage. Top panels: left Eph1; center Otx; right FGF9/16/20. The FGF9/16/20 Pm04 enhancer has so far only been tested in *Ciona intestinalis*. Bottom panels: schematic representation of the activity of the 6 enhancers. The left half corresponds to the more distal enhancer (violet), the right half to the more proximal one. Light blue: shared activity, violet: distal enhancer-specific activity, orange: proximal enhancer-specific activity.

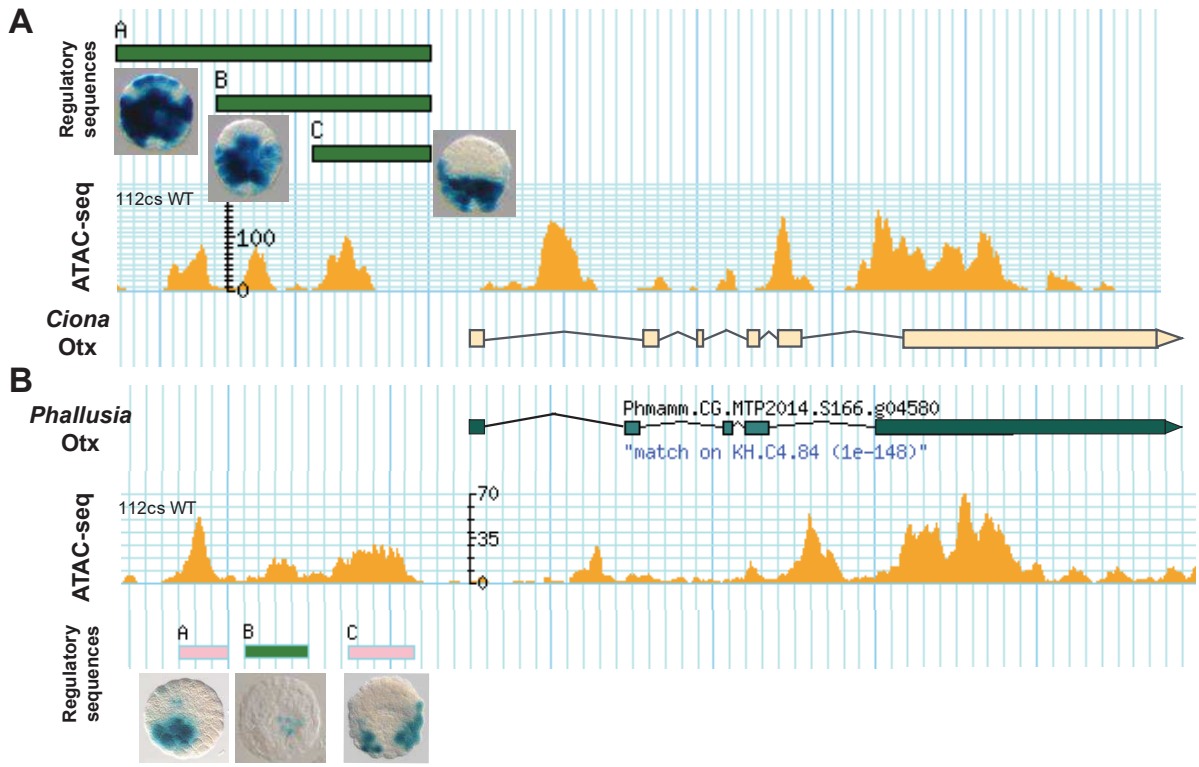


Figure 4: Similarity of the chromatin accessibility landscapes at the *Otx* locus in *Ciona* and *Phallusia.* A) *Ciona* landscape, showing ATAC-seq landscape, tested regulatory regions and their pattern of reporter gene activity at the early gastrula stage. B) *Phallusia* landscape, with tested regulatory regions and their pattern of reporter gene activity at the early gastrula stage.

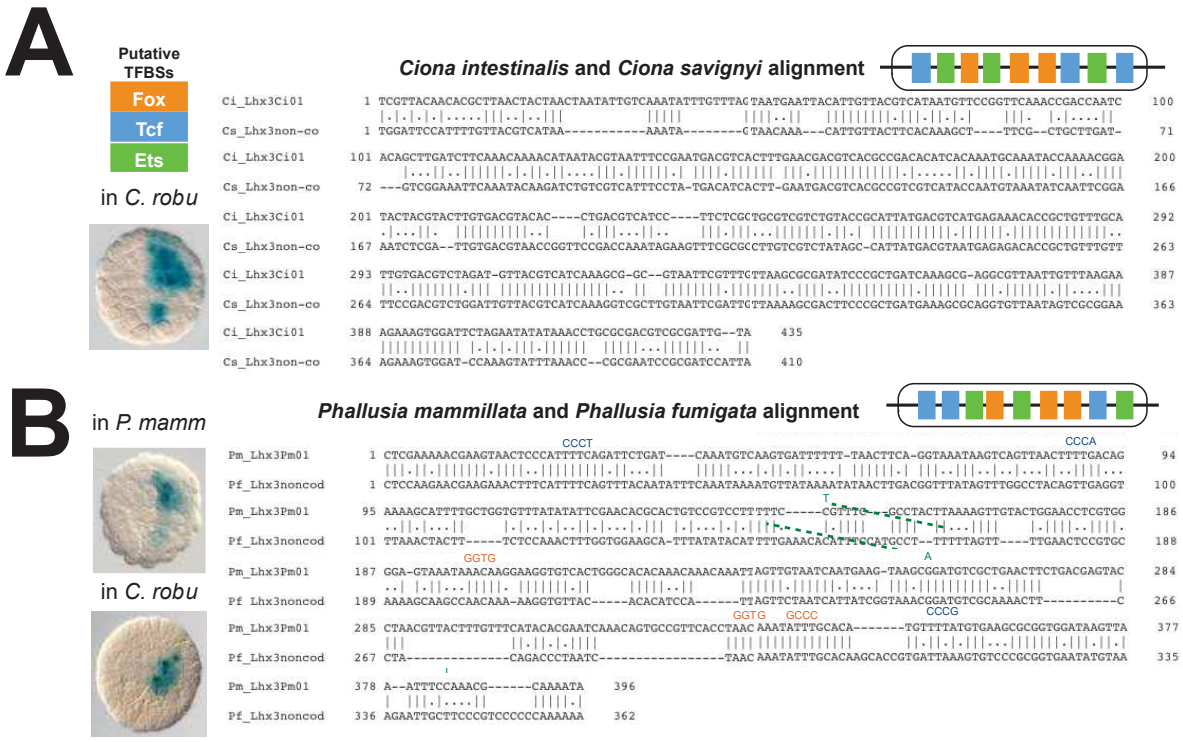


Figure 5: Conserved regulatory syntax of the *Ciona* and *Phallusia* early endodermal enhancers of the *Lhx3/4* gene. XXXX

SUPPLEMENTARY FIGURE

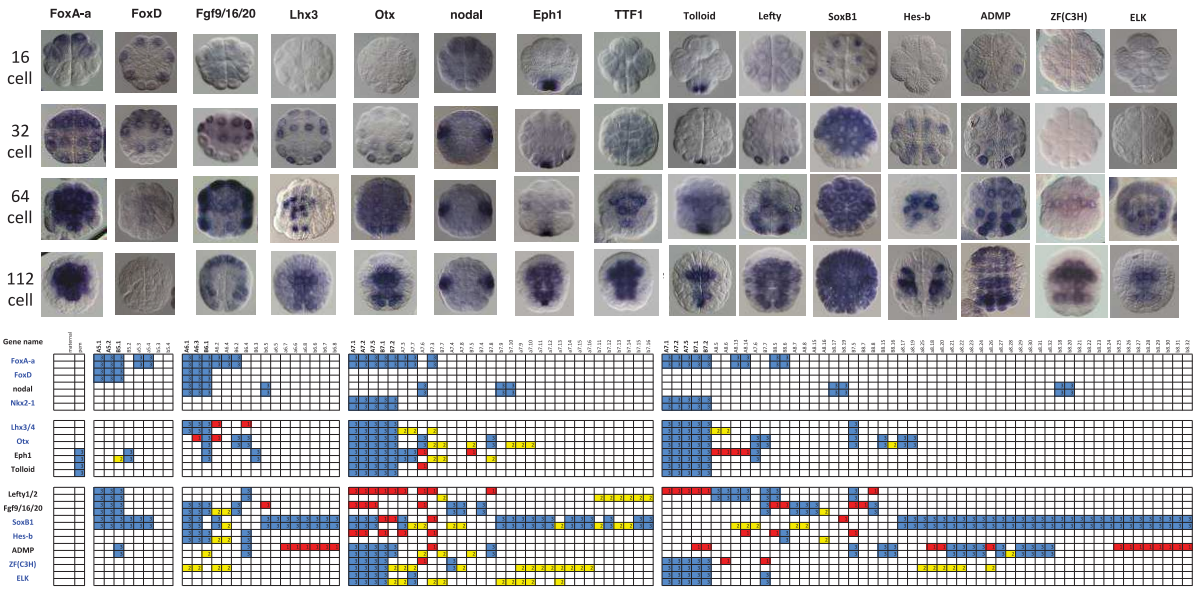


Figure S1: Qualitative expression profiles of 15 endodermal regulatory genes up to the gastrula stage. A) Representative pictures of the expression profiles of the genes at the 4 stages. B) Summary table of the pregastrula expression profiles of 15 endodermal regulatory genes in *Ciona* and *Phallusia*. For each gene in the left column (blue: transcription factors, black: signalling molecule) the table indicates for each cell up to the 112-cell stage (top row, in bold endoderm precursors) its expression in both species (light blue), in *Ciona intestinalis* only (yellow) and in *Phallusia mammillata* only (red). Expression for each gene is encoded in two rows: the top row presents the *Phallusia* expression, the bottom row the *Ciona* expression. Genes are organized in three blocks: top, identical expression; middle very similar expression; bottom, more divergent expression.

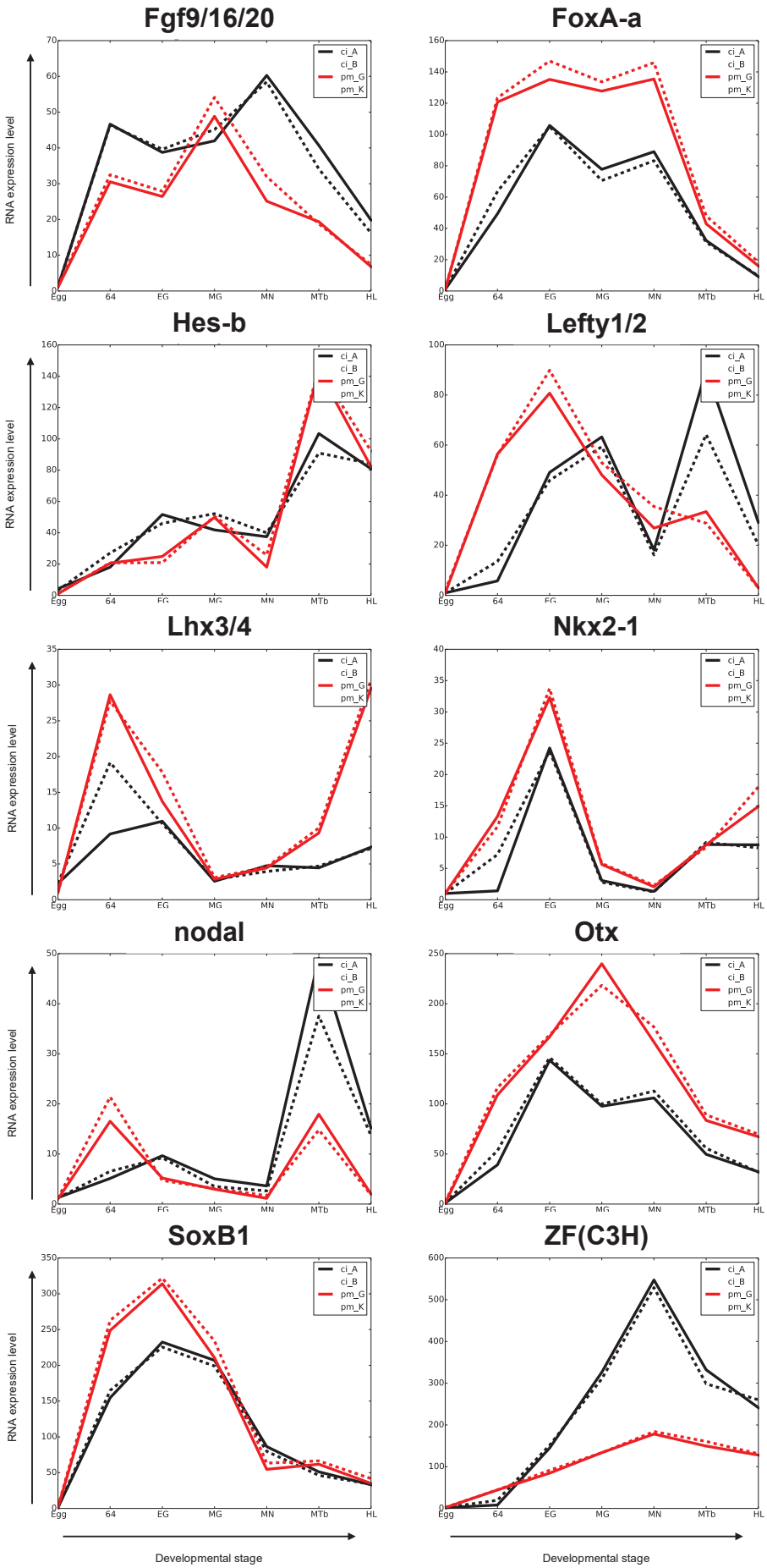

Fgf9/16/20
FoxA-a
Hes-b
Lefty1/2
Lhx3/4
Nkx2-1
nodal
Otx
SoxB1
ZF(C3H)
RNA expression level
ci_A
ci_B
pm_G
pm_K
Egg
64
EG
MG
MN
MTb
HL
Developmental stage

Figure S2: Semi quantitative expression patterns by RNA-seq of endodermal genes in whole embryos during development. Expression was determined by RNA-seq at the same 7 stages in *Phallusia mammillata* and *Ciona intestinalis*. Y-axis: FPKM values, X-axis: developmental stages (64 : 64-cell stage; EG: early gastrula, stage 11; MG, mid-gastrula stage 12; MN: mid neurula, stage 15; MTb: mid tailbud, stage 21; HL: Hatching larvae stage 26. Red: *Phallusia mammillata*. Black: *Ciona intestinalis* sequences, aligned onto the *Ciona robusta* genome. Dotted line: average of the two replica.

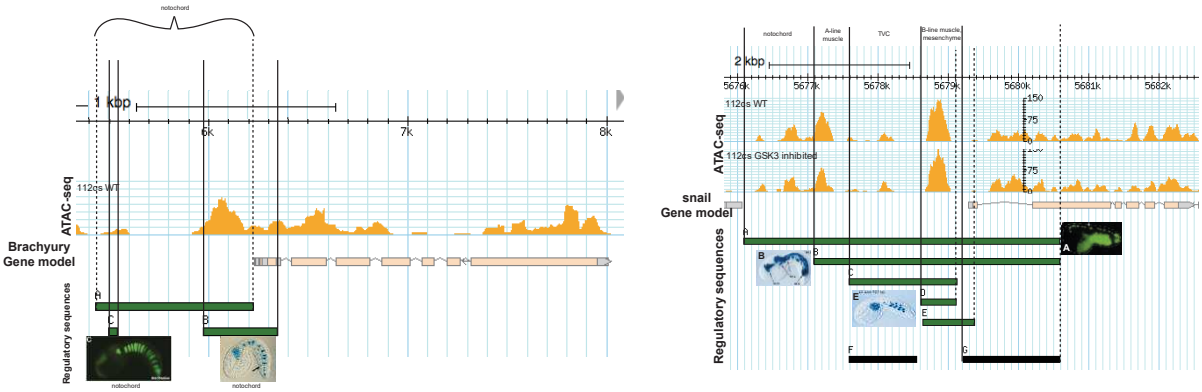


Figure S3: ATAC-seq profiles at the *Brachyury* and *Snail* loci. Chromatin accessibility (orange) is compared to the position of active cis-regulatory sequences. Note that Brachyury sequences C and B both drive expression in the notochord, though their precise timing of onset of expression is unknown.

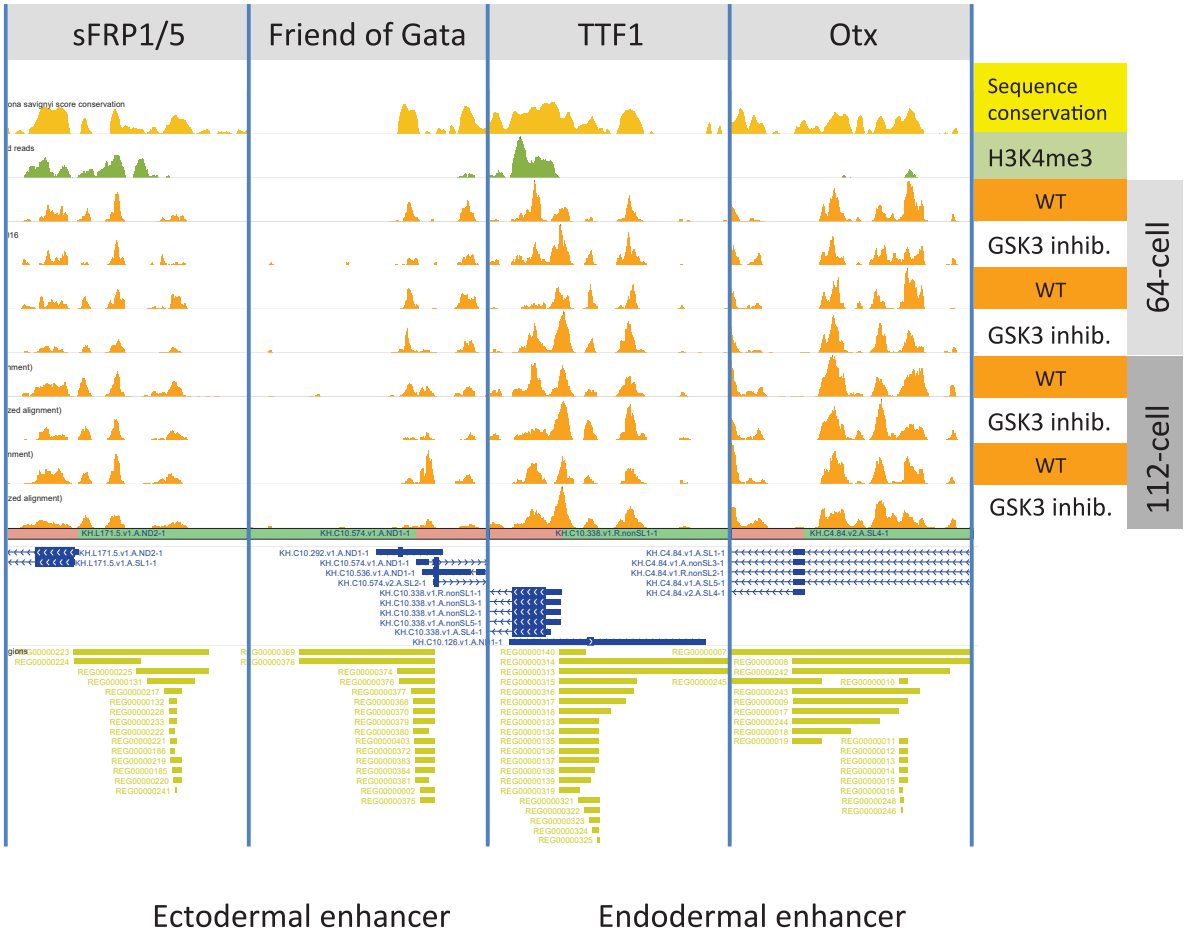
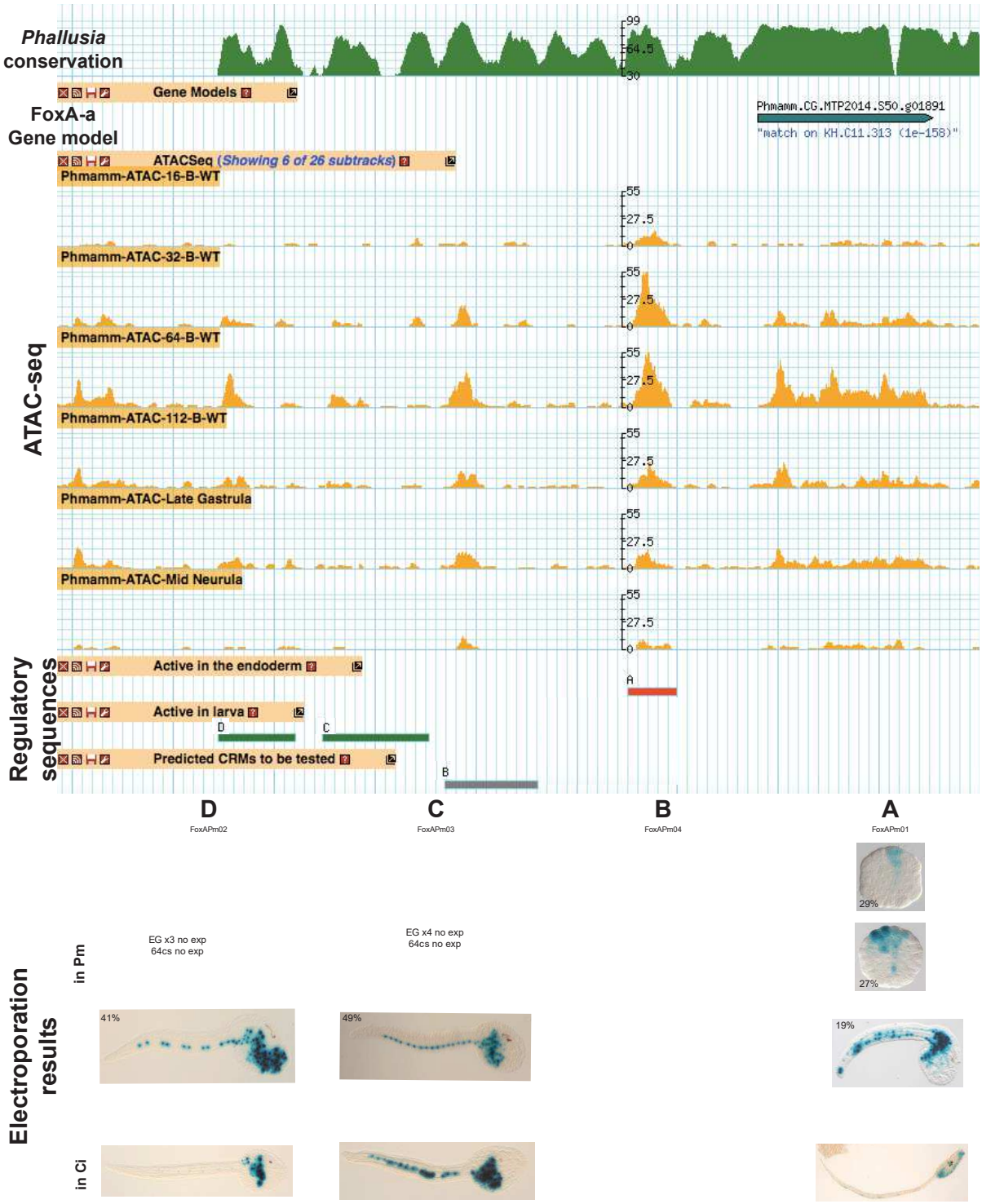
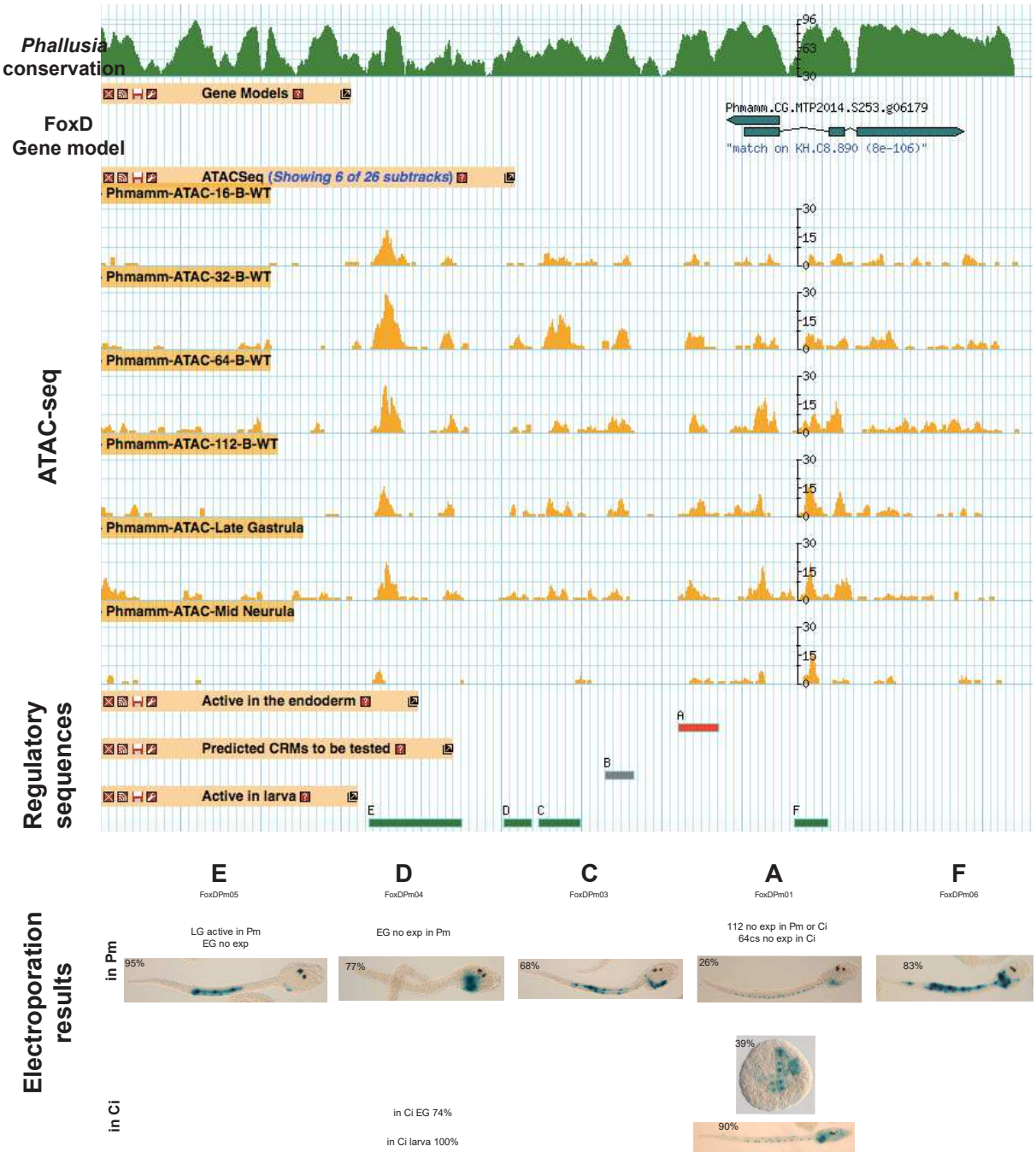


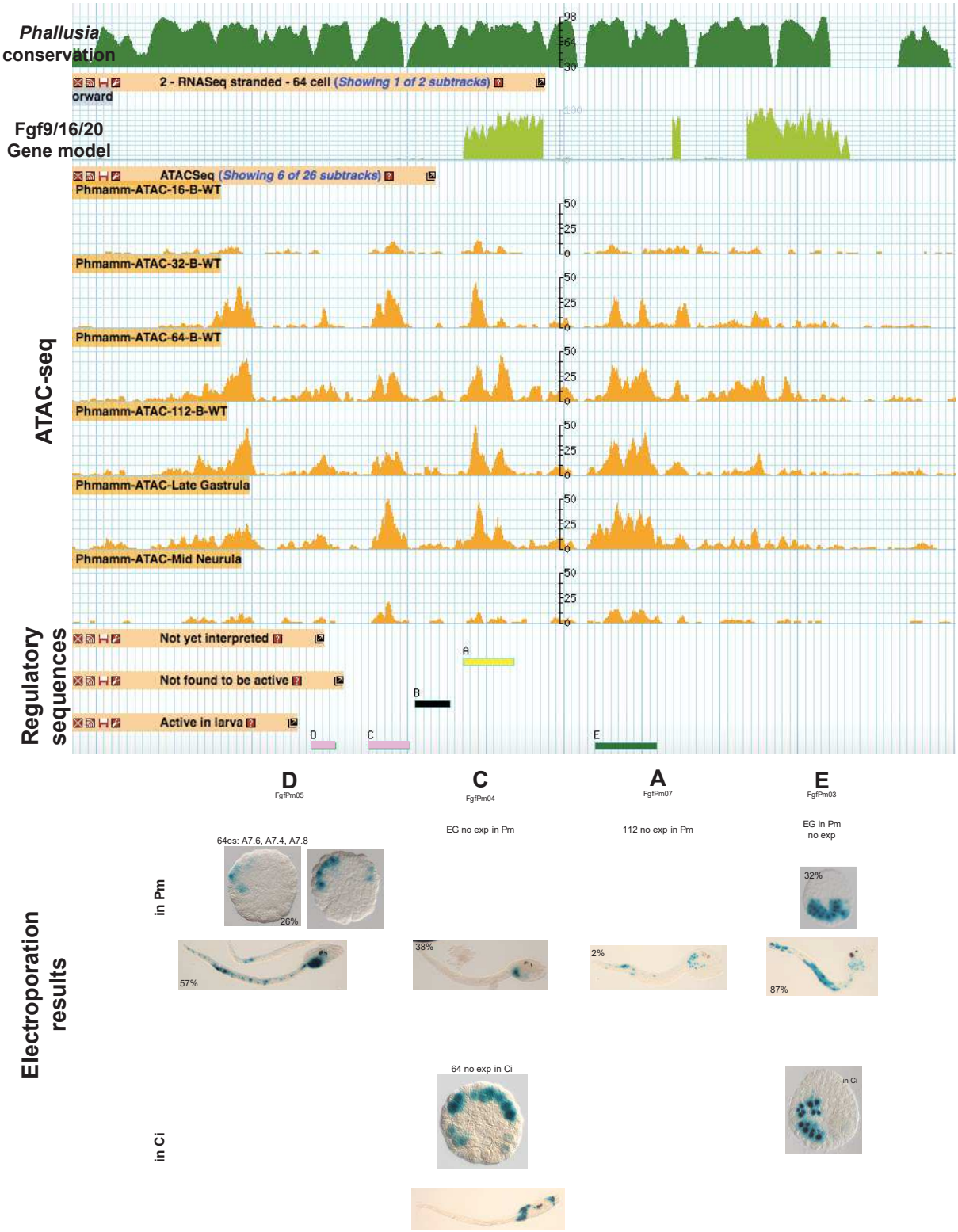
Figure S4: Chromatin accessibility landscapes at 4 *Ciona intestinalis* loci: Reproducibility, stability and effect of GSK3 inhibition. The figure compares ATAC-seq signals to ChIP-seq data for the promoter mark H3K4me3, to sequence conservation between *Ciona robusta* and *Ciona savignyi*, and to the position of known endodermal (pink) and ectodermal (light blue) *Ciona* enhancers.


Phallusia conservation
99
64.5
30
Gene Models
FoxA-a Gene model
Phmamm.CG.MTP2014.S50.g01891
"match on KH.C11.313 (1e-158)"
ATACSeq (Showing 6 of 26 subtracks)
ATAC-seq
Phmamm-ATAC-16-B-WT
Phmamm-ATAC-32-B-WT
Phmamm-ATAC-64-B-WT
Phmamm-ATAC-112-B-WT
Phmamm-ATAC-Late Gastrula
Phmamm-ATAC-Mid Neurula
55
27.5
0
Regulatory sequences
Active in the endoderm
Active in larva
Predicted CRMs to be tested
A
B
C
D
D
FoxAPm02
C
FoxAPm03
B
FoxAPm04
A
FoxAPm01
Electroporation results
in Pm
in Ci
EG x3 no exp
64cs no exp
EG x4 no exp
64cs no exp
41%
49%
29%
27%
19%

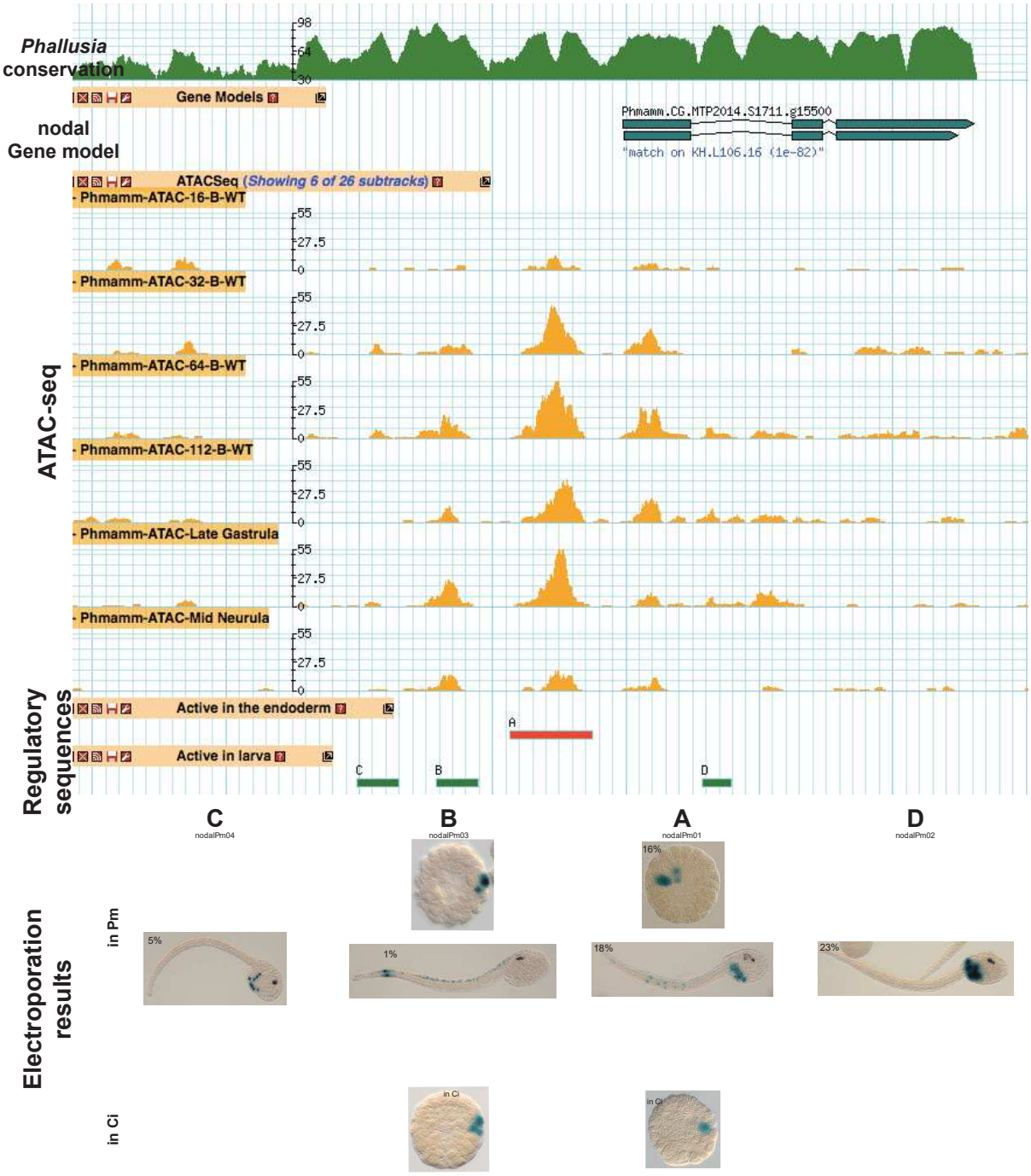
A) *FoxAa*


Phallusia conservation
96
63
30
Gene Models
FoxD Gene model
Phmamm.CG.MTP2014.S253.g06179
"match on KH.C8.890 (8e-106)"
ATACSeq (Showing 6 of 26 subtracks)
ATAC-seq
Phmamm-ATAC-16-B-WT
Phmamm-ATAC-32-B-WT
Phmamm-ATAC-64-B-WT
Phmamm-ATAC-112-B-WT
Phmamm-ATAC-Late Gastrula
Phmamm-ATAC-Mid Neurula
30
15
0
Regulatory sequences
Active in the endoderm
Predicted CRMs to be tested
Active in larva
A
B
C
D
E
F
Electroporation results
E
FoxDPm05
D
FoxDPm04
C
FoxDPm03
A
FoxDPm01
F
FoxDPm06
LG active in Pm
EG no exp
EG no exp in Pm
112 no exp in Pm or Ci
64cs no exp in Ci
in Pm
95%
77%
68%
26%
83%
39%
in Ci
in Ci EG 74%
in Ci larva 100%
90%

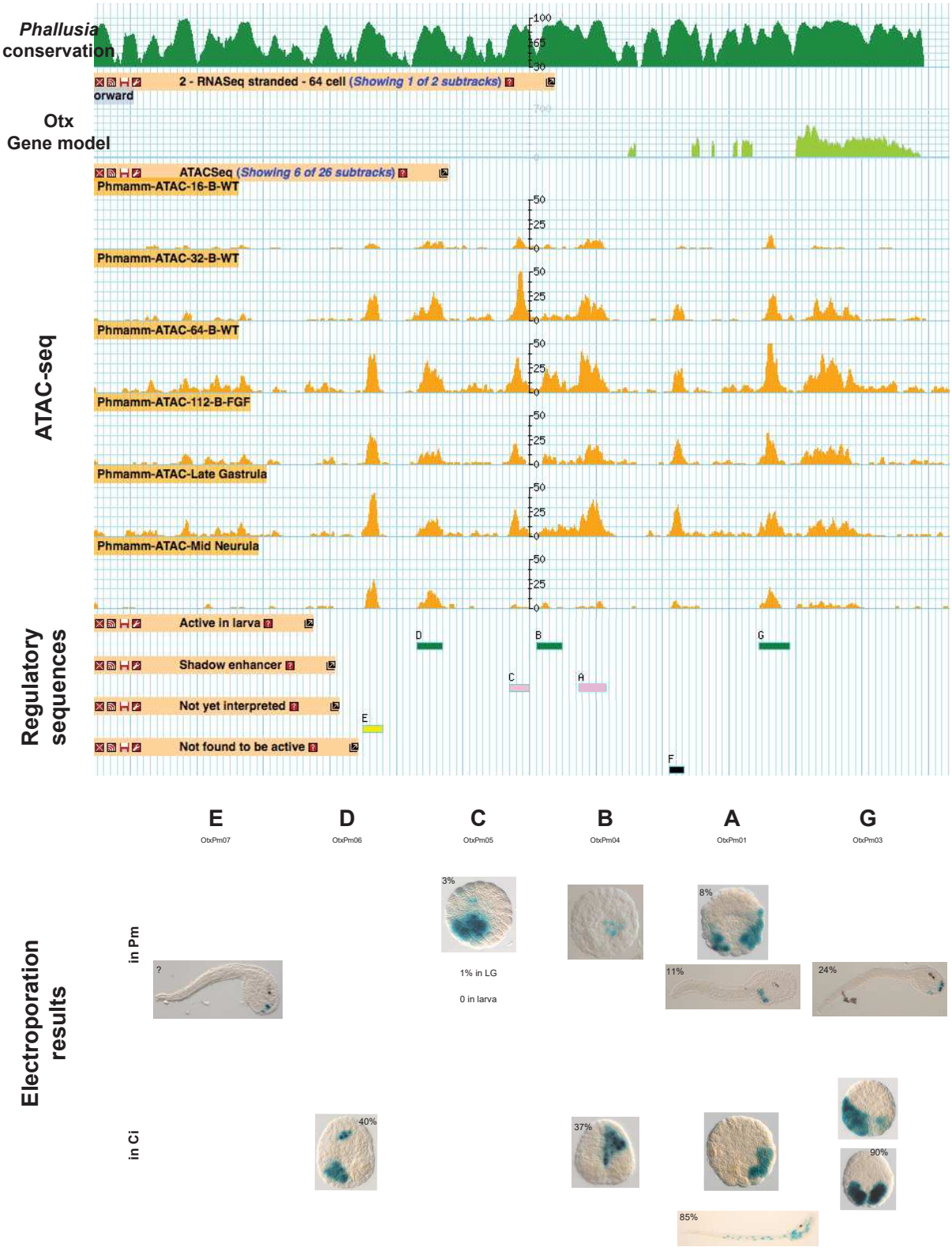
B) *FoxD*


Phallusia conservation
2 - RNASeq stranded - 64 cell (Showing 1 of 2 subtracks)
orward
Fgf9/16/20 Gene model
ATACSeq (Showing 6 of 26 subtracks)
Phmamm-ATAC-16-B-WT
Phmamm-ATAC-32-B-WT
Phmamm-ATAC-64-B-WT
Phmamm-ATAC-112-B-WT
Phmamm-ATAC-Late Gastrula
Phmamm-ATAC-Mid Neurula
ATAC-seq
Not yet interpreted
Not found to be active
Active in larva
Regulatory sequences
D
FgfPm05
C
FgfPm04
A
FgfPm07
E
FgfPm03
64cs: A7.6, A7.4, A7.8
EG no exp in Pm
112 no exp in Pm
EG in Pm no exp
in Pm
26%
32%
57%
38%
2%
87%
64 no exp in Ci
in Ci
Electroporation results

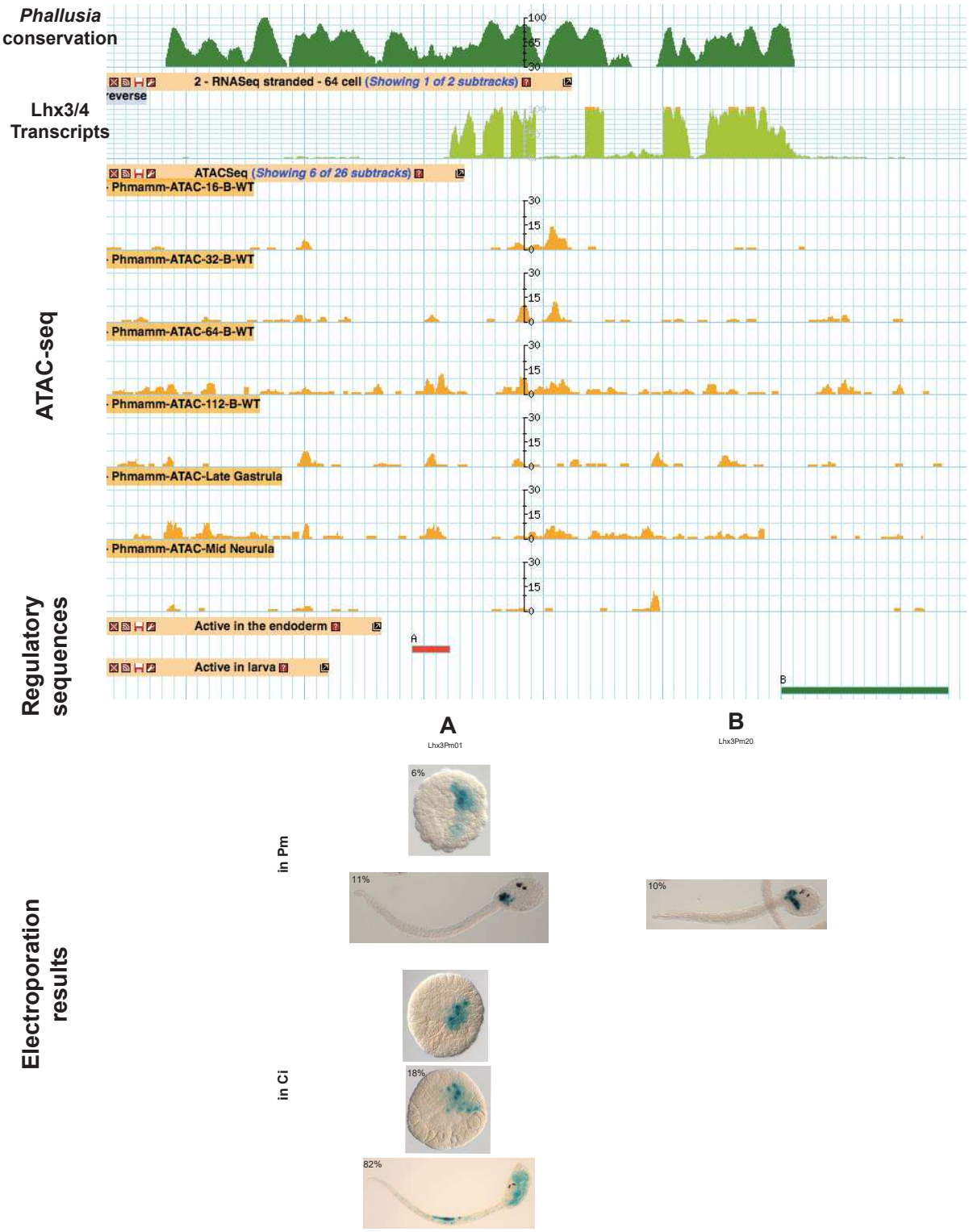
C) *FGF9/16/20*


Phallusia conservation
98
64
30
Gene Models
nodal Gene model
Phmamm.CG.MTP2014.S1711.g15500
"match on KH.L106.16 (1e-82)"
ATAC-seq
ATACSeq (Showing 6 of 26 subtracks)
Phmamm-ATAC-16-B-WT
Phmamm-ATAC-32-B-WT
Phmamm-ATAC-64-B-WT
Phmamm-ATAC-112-B-WT
Phmamm-ATAC-Late Gastrula
Phmamm-ATAC-Mid Neurula
55
27.5
0
Regulatory sequences
Active in the endoderm
Active in larva
A
B
C
D
Electroporation results
in Pm
in Ci
C
nodalPm04
B
nodalPm03
A
nodalPm01
D
nodalPm02
5%
1%
16%
18%
23%

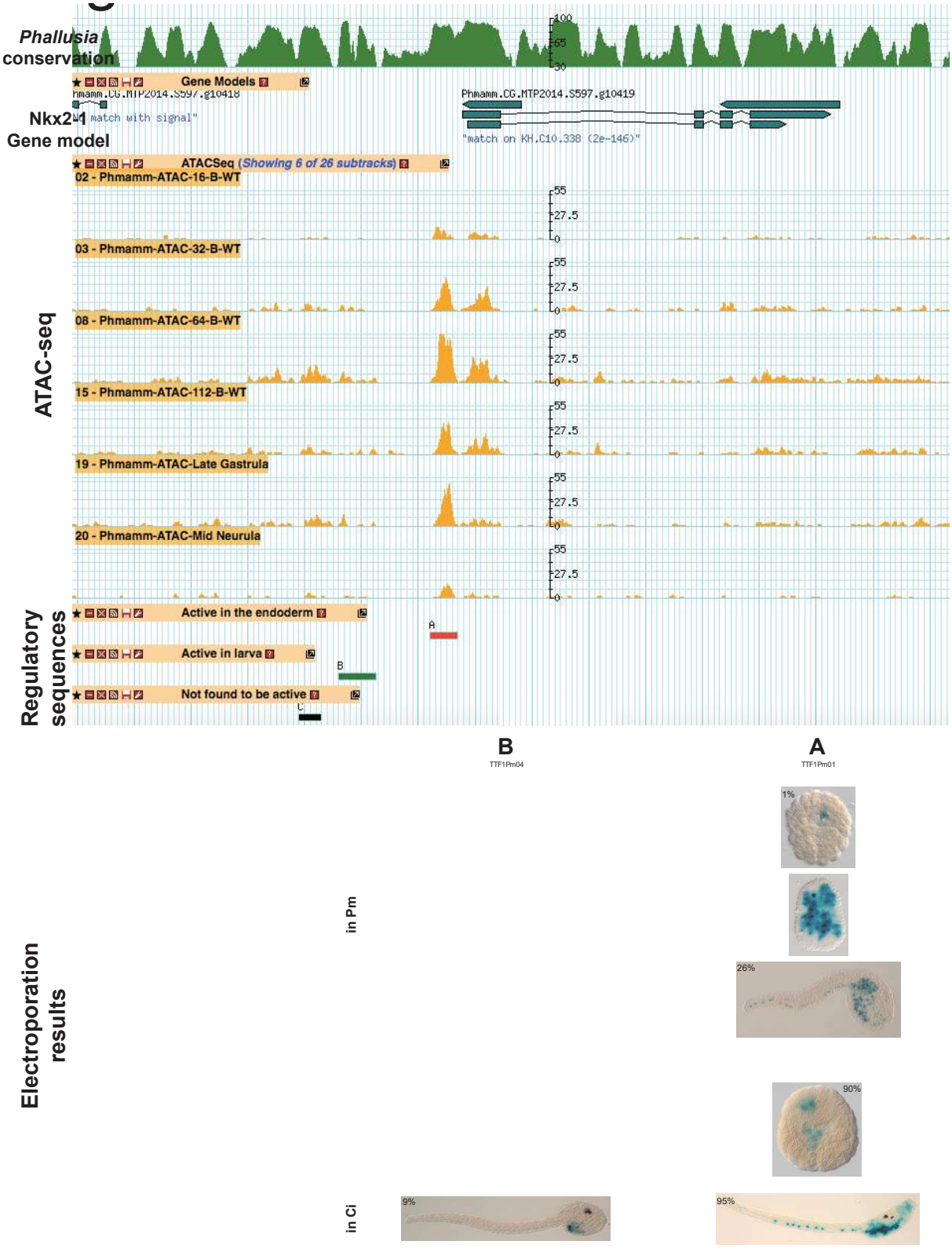
D) *Nodal*


Phallusia conservation
2 - RNASeq stranded - 64 cell (Showing 1 of 2 subtracks)
orward
Otx Gene model
ATACSeq (Showing 6 of 26 subtracks)
Phmamm-ATAC-16-B-WT
Phmamm-ATAC-32-B-WT
Phmamm-ATAC-64-B-WT
Phmamm-ATAC-112-B-FGF
Phmamm-ATAC-Late Gastrula
Phmamm-ATAC-Mid Neurula
ATAC-seq
Active in larva
Shadow enhancer
Not yet interpreted
Not found to be active
Regulatory sequences
E
D
C
B
A
G
OtxPm07
OtxPm06
OtxPm05
OtxPm04
OtxPm01
OtxPm03
in Pm
in Ci
3%
8%
1% in LG
0 in larva
11%
24%
40%
37%
90%
85%
Electroporation results

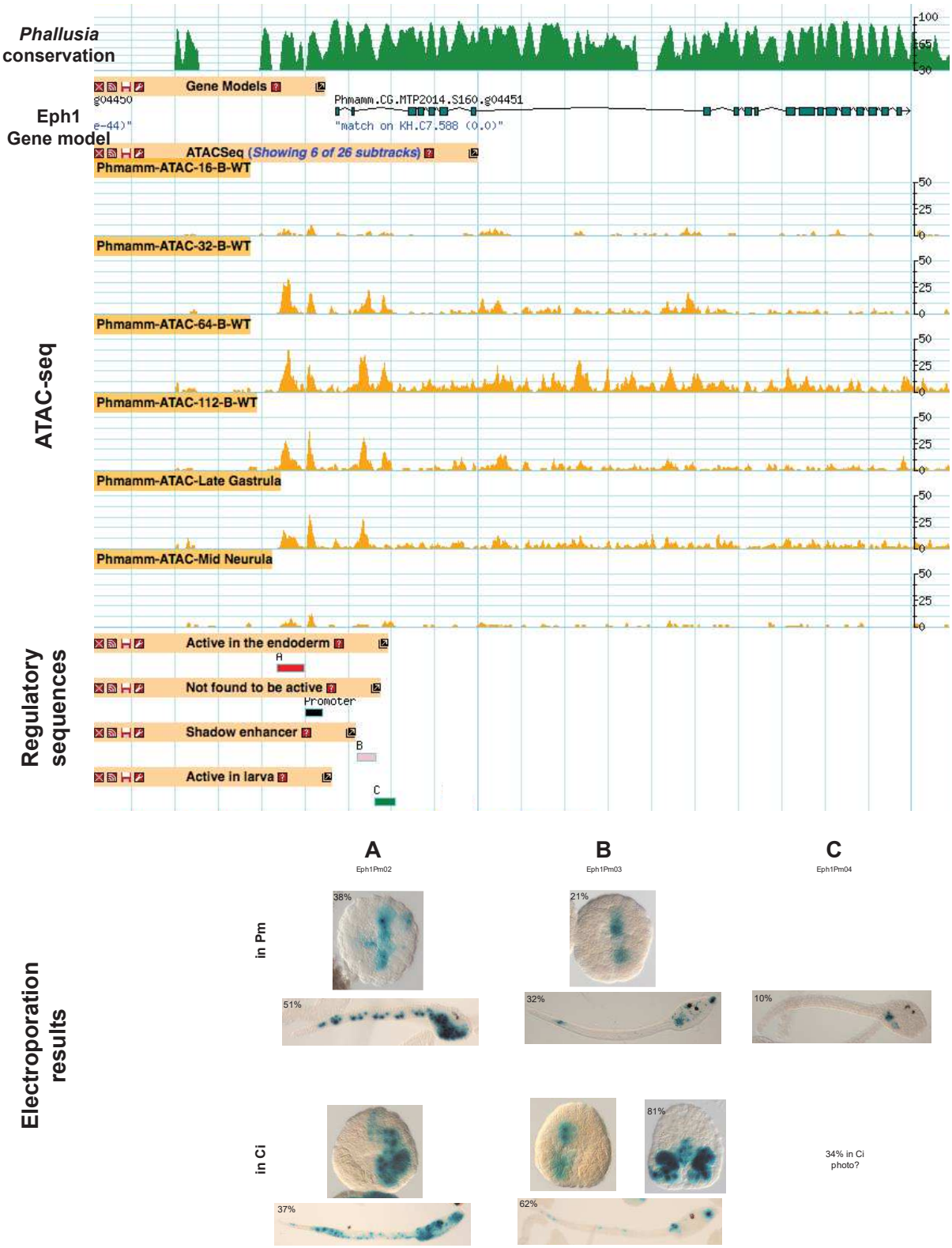
E) *Otx*


Phallusia conservation
2 - RNASeq stranded - 64 cell (Showing 1 of 2 subtracks)
Lhx3/4 Transcripts
ATACSeq (Showing 6 of 26 subtracks)
Phmamm-ATAC-16-B-WT
Phmamm-ATAC-32-B-WT
Phmamm-ATAC-64-B-WT
Phmamm-ATAC-112-B-WT
Phmamm-ATAC-Late Gastrula
Phmamm-ATAC-Mid Neurula
ATAC-seq
Active in the endoderm
Active in larva
Regulatory sequences
A
Lhx3Pm01
B
Lhx3Pm20
in Pm
6%
11%
10%
in Ci
18%
82%
Electroporation results

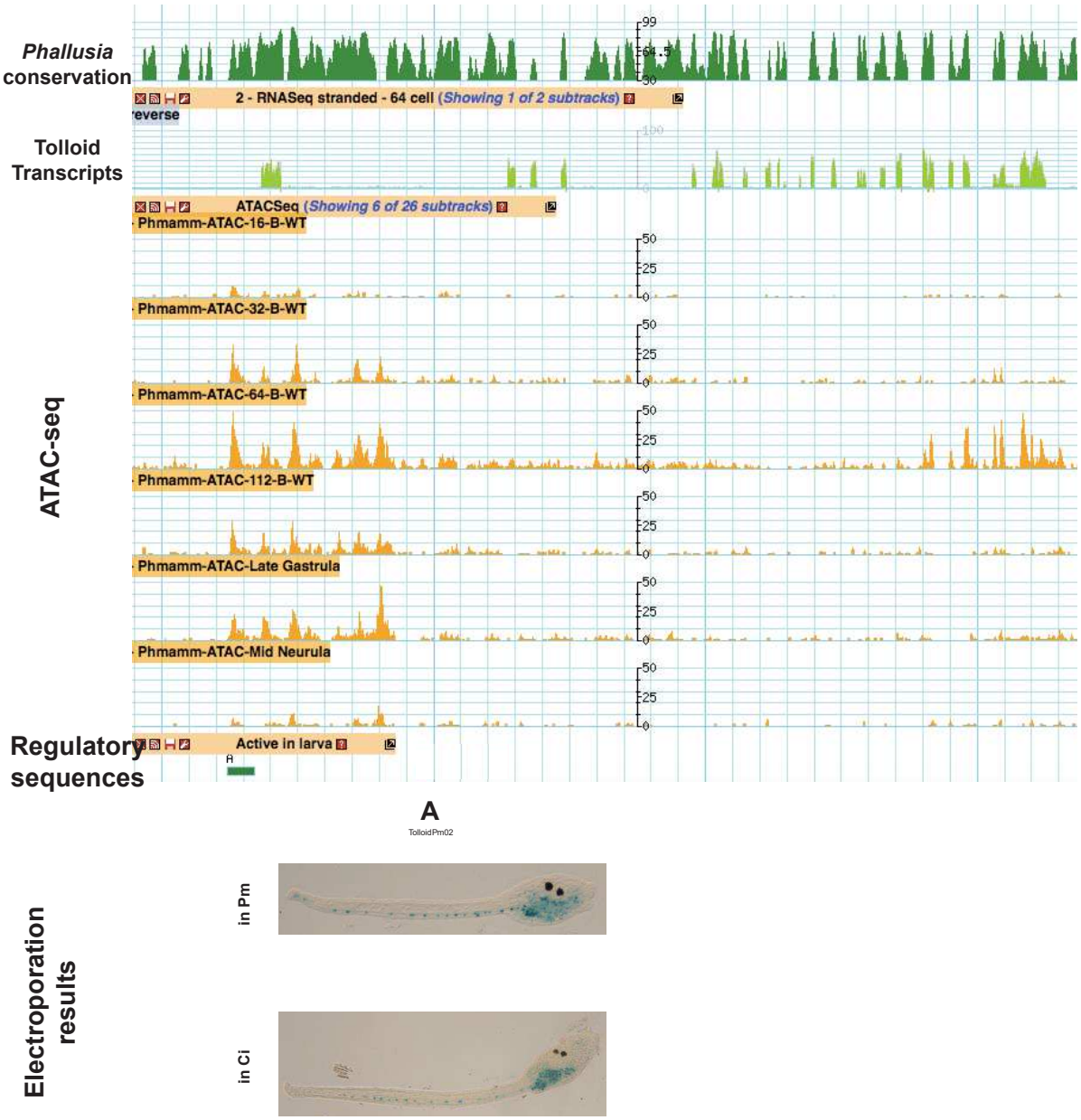
F) *Lhx3/4*


Phallusia conservation
Gene Models
Nkx2-1 Gene model
"match with signal"
Phmamm.CG.MTP2014.S597.g10419
"match on KH.C10.338 (2e-146)"
ATACSeq (Showing 6 of 26 subtracks)
02 - Phmamm-ATAC-16-B-WT
03 - Phmamm-ATAC-32-B-WT
08 - Phmamm-ATAC-64-B-WT
15 - Phmamm-ATAC-112-B-WT
19 - Phmamm-ATAC-Late Gastrula
20 - Phmamm-ATAC-Mid Neurula
ATAC-seq
Active in the endoderm
Active in larva
Not found to be active
Regulatory sequences
B
TTF1Pm04
A
TTF1Pm01
Electroporation results
in Pm
in Ci
1%
26%
90%
9%
95%

G) *TTF1*


Phallusia conservation
Eph1 Gene model
Gene Models
Phmamm.CG.MTP2014.S160.g04451
"match on KH.C7.588 (0.0)"
ATACSeq (Showing 6 of 26 subtracks)
Phmamm-ATAC-16-B-WT
Phmamm-ATAC-32-B-WT
Phmamm-ATAC-64-B-WT
Phmamm-ATAC-112-B-WT
Phmamm-ATAC-Late Gastrula
Phmamm-ATAC-Mid Neurula
ATAC-seq
Active in the endoderm
Not found to be active
Promoter
Shadow enhancer
Active in larva
Regulatory sequences
A
B
C
Eph1Pm02
Eph1Pm03
Eph1Pm04
in Pm
in Ci
38%
21%
51%
32%
10%
81%
37%
62%
34% in Ci photo?
Electroporation results

H) *Eph1*



I) *Tolloid*

Figure S5: Summary of the chromatin accessibility landscape, tested candidate cis-regulatory regions and their activity at the *Phallusia mammillata* loci of the 10 genes studied. A) *FoxAa*; B) *FoxD*; C) *FGF9/16/20*; D) *Nodal*; E) *Otx*; F) *Lhx3/4*; G) *TTF1*; H) *Eph1*; I) *Tolloid*; Each panel shows from top to bottom: *Phallusia mammillata/fumigata* sequence conservation; *Phallusia* gene model or RNA-seq if the model is inaccurate; ATAC seq profiles at successive stages; regions tested for *cis*-regulatory activity with their activity.

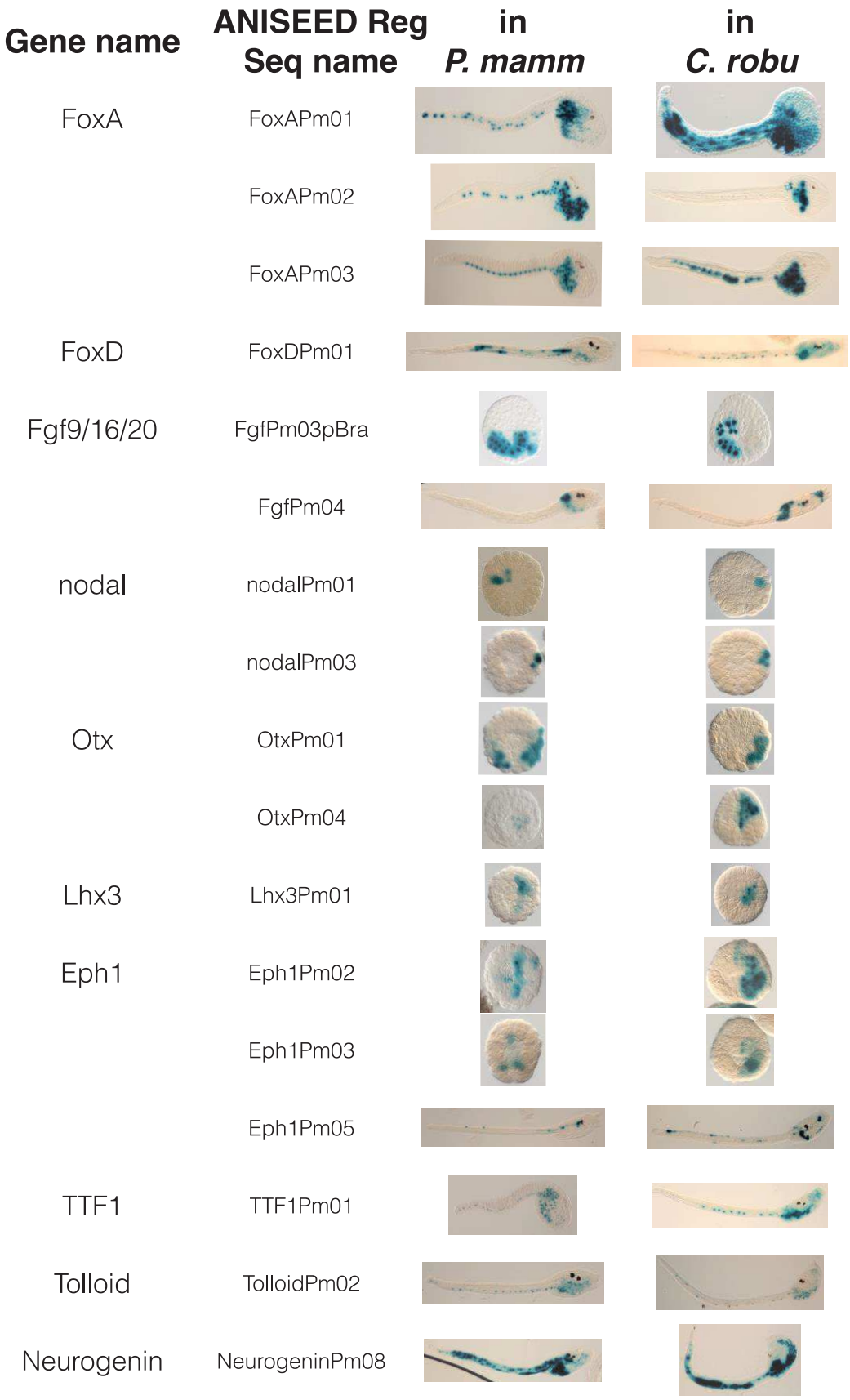


Figure S6: Distribution of ATAC-seq peaks length and inter-peak distances.

Discussion

I. Analysing and interpreting WT expression patterns

Previous work done in the lab, by Jacques Piette and Christelle Dantec, comparing *Ciona intestinalis* and *Phallusia mammillata* whole embryo RNA temporal expression levels found that there seemed to be significant differences in expression between these two species. Following from this work, the starting point of my project was to compare qualitative expression patterns at the cellular level in *Phallusia mammillata* to the known *Ciona intestinalis* expression patterns at several stages leading up to gastrulation.

1. The limitations of fixed samples

The major problem was comparing the *Phallusia* expression patterns to the published *Ciona* expression data. First of all, these data have not been generated by the same person. Therefore, the experiments have not been carried out in the exact same way, for example with regards to the staining time.

Secondly, an issue is comparing the expression at the equivalent times in development. Expression in ascidians can be very dynamic; RNA expression can be very different in the cells between the beginning and the end of the same developmental stage. Different works may decide to focus on just one time point within a stage as the reference stage. What I have done is try to regroup all of the expression within a stage into the same stage. This may not have been kept into account for the *Ciona* work.

Due to these different variables between expression data, the profiles may not have necessarily been interpreted in the same manner. Therefore, differences in expression data between the animals are either due to differences in experiments, in interpretation of the data, in slightly different developmental timings or they could be actual differences in expression between the two species.

2. Active transcription: nascent RNA

For my project, to determine the GRNs, I needed to follow gene expression for two reasons. Firstly, I wanted to know specifically in which cells the TFs and signalling molecules were available to activate downstream genes. Secondly, I wanted to know exactly at what stage these genes were activated. When the mother cell had already expressed the gene in question, it is not necessarily possible to determine if the gene is still active in the daughter cells because the RNA is often still present.

To by pass this question, it would have been possible to perform ISH specifically for nascent RNA. Detecting nascent RNA by designing the probes to bind to the intron before it has been spliced out would show if the gene is still under transcription and therefore still being regulated. We considered this doing these experiments, however, we thought that the signal would be so weak that it was not worth doing the experiments. This is not the most interesting information to generate; quantifying the expression would have been more fruitful.

3. Quantitative expression profiles

Although the colourimetric ISH results are easily interpretable, they are purely qualitative and are not as sensitive as other techniques. With regards to quantitative data, we considered two possibilities: single-molecule FISH (smFISH) (Femino et al., 1998) and RNA-sequencing either in whole embryo at each of the stages of interest for the early GRNs. Ideally, single cell RNA-seq would elucidate expression levels genome-wide, however, this method would be quite laborious at each developmental stage. Despite this, the advantages would be quite considerable by deep sequencing. This could potentially inform us when RNA levels are very low, perhaps too low to be detected by other techniques.

Studying expression levels variability within and between species is particularly interesting to analyse the robustness of expression levels. The whole-embryo RNA-seq that was performed in the lab already showed differences between *Phallusia* and *Ciona* and even within a species for certain genes. It would be very interesting to analyse what are the parameters contributing to the variability in expression levels (for example, differences in the *cis*-regulatory sequences such as number of TFBSs or the effect of spacer sequences). What has caused these differences in expression levels? Are expression levels even important for ascidian development?

II. A different, more classical, approach to the endoderm GRNs

The bulk of my project was more oriented towards descriptive work based on GRNs that had already been studied in *Ciona*. It would have been interesting to add a more functional aspect to my work by expanding the GRNs driving the endodermal fate.

1. Classical approach

As described in the introduction of this thesis, Lhx3/4 and Nkx2-1 TFs are known to be involved in endodermal fate determination. It would have been interesting to identify the genes immediately downstream in the regulatory cascade. One approach could have

been to compare expression between wild type and morpholino embryos or mRNA-injected embryos for Lhx3/4 and Nkx2-1.

Lhx3/4 MO embryos completely inhibit endodermal development; however, in Nkx2-1 MO, the endoderm is only down regulated. This could mean that downstream genes involved in endodermal fate would potentially have the same profile: completely down regulated in Lhx3/4 MO and only partly down regulated in Nkx2-1 MO embryos. The combination of RNA-seq data of these 3 conditions could provide the possibility of making a list of candidate genes to further analyse. Although, a more reliable tool to knock-out gene expression could also be to remove or inactivate a gene by CRISPR; recent works using CRISPR to disable a gene that had previously been found to have a certain function by other techniques have found that the older techniques did not tell the whole story (Ledford).

2. New take on a classical idea

A newer, more exciting tool for this analysis work would be to target gene expression in a cell or tissue-specific manner by CRISPR. Following the enhancer screen in this project, a next step would to be to remove or partially remove an enhancer driving gene expression in a specific tissue. Preliminary analysis of the enhancer activities found that most of the enhancers seem to be used only once during early embryogenesis. Targeting these modular-style enhancers would mean that we could determine the importance of a gene in a specific tissue without necessarily disrupting the development of the rest of the embryo.

III. Gene regulation and chromatin topology: interpreting the ATAC-seq data

The collaborative ATAC-seq work with José Luis Gomez-Skarmeta's lab was a turning point in my project. It was not only a great experience to work side by side with Marta Silvia Magri but also the ATAC-seq actual brought a wealth of information about nucleosome positioning and regulation in ascidians.

The ATAC-seq data revealed an unanticipated behaviour of ascidian chromatin through time. It was expected that only currently-active regulatory sequences would be open and that re-positioning of nucleosomes would be specific to enhancer activity. We expected to be able to identify regulatory sequences actively driving transcription by their absence of nucleosomes. However, almost all of the 36 enhancers that we found are located in an open region from the 16 cell-stage until mid neurula. This raised many questions regarding the function of chromatin topology and its effect on gene expression. Why are these regulatory regions open far in advance of gene activation? Why are there so many open regulatory sequences?

1. Do maternal factors control chromatin opening of most enhancers genome wide?

Of the genes studied, almost all of the regulatory sequences that were open at 16 cell-stage remain open until the mid neurula stage. In many cases, these regulatory

sequences are active much later in development; this raises the question: what is causing these regions to be open?

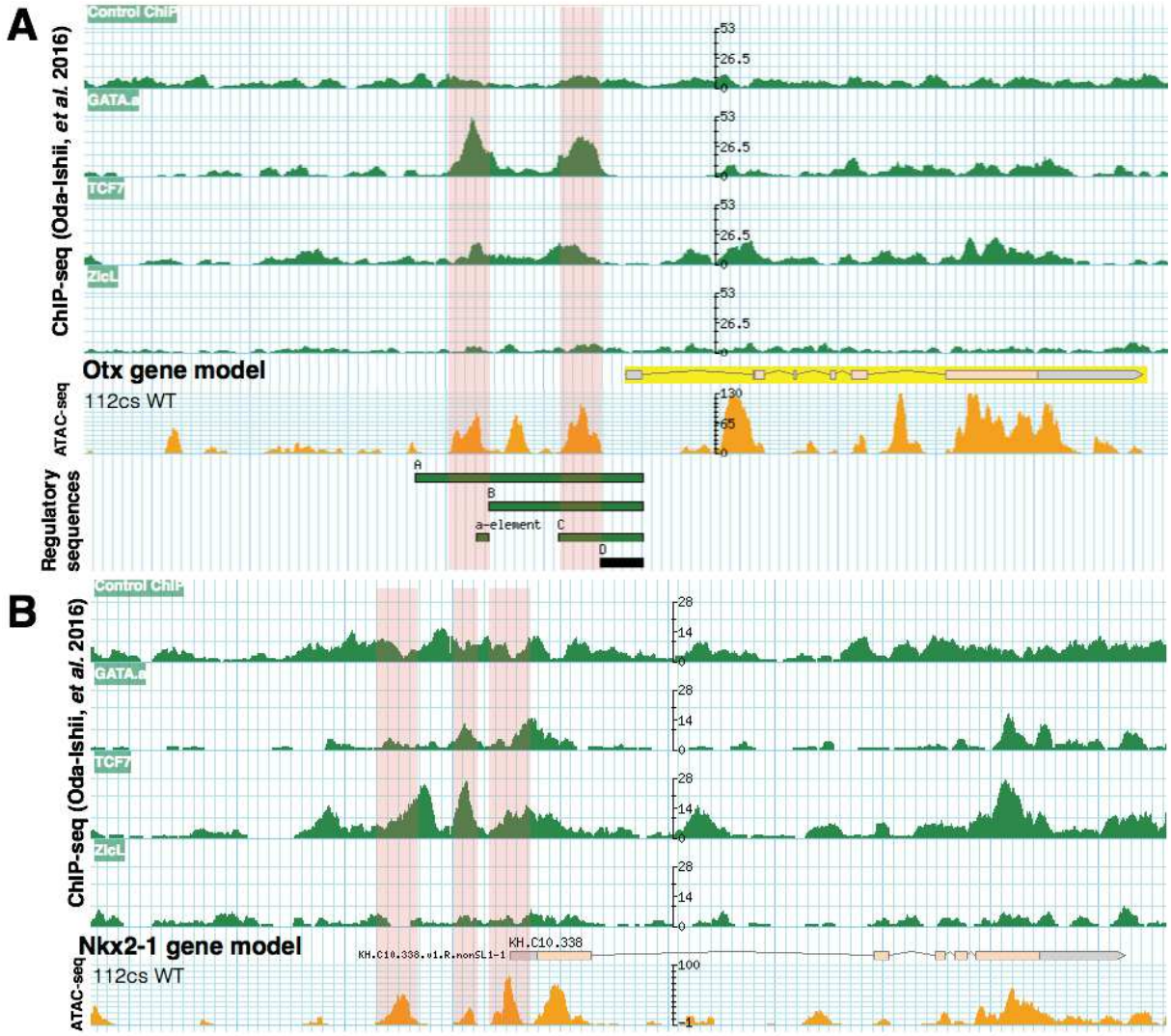


Figure 20. There is an enrichement of maternal factors binding within the open chromatin, as seen around A) Otx and B) Nkx2-1. (Adapted from Oda-Ishii et al., 2016)

This could potentially be due to the initial binding of pioneer TFs by as early as the 16-cell stage. Maternal β-catenin/Tcf is known to be an important driver of vegetal expression to establish, along with Gata.a and Zic-r.a, the first zygotic expression patterns. Many of the accessible chromatin regions in Ciona are enriched for Tcf binding (e.g. Otx Figure 20A), even before gene expression has initiated (e.g. Nkx2-1 Figure 20B). What is remarkable is that the maternal TFs are bound many stages before the activation and expression of the genes.

The onset of zygotic gene expression is thought to be around the 8- to 16-cell stage. TF FoxA-a is thought to be one of the first zygotically expressed genes; FoxA gene is a known pioneer TF in mouse embryonic endoderm and is responsible for nucleosome displacement (Gualdi et al., 1996; Iwafuchi-Doi et al., 2016). This could mean that by the

16-cell stage, zygotic factors may already be binding to the DNA. In this case, the open chromatin at 16-cell stage represents regions that are bound by both maternal and zygotic factors. To appreciate the extent of genes that are partially activated by maternal factors, we would need to perfom ATAC-seq at 8cs as well. It could also be interesting to perform ATAC-seq in β-catenin inhibited embryos in the earlier stages.

2. Do these enhancers work alone?
As previously mentioned, we had not anticipated finding so many open regions that are open throughout early development. Finding that shadow enhancers are more common in ascidians than expected does not fully explain the extent of open regions. What is the purpose of having so many open regulatory regions at the same time, from the start of development?

One possibility that was proposed in the previous section is that transcription factor availability is perhaps the main limiting factor for activating gene expression. However, perhaps enhancer competition is also a limiting factor or, on the contrary, several enhancers work together at a time to drive expression. My project has not addressed these questions. Further tests would have to be performed to observe the effect of removing each enhancer. Testing individual enhancers does not necessarily represent the true WT activity – it just represents the potential activity of the enhancer.

IV. Transcription factor binding sites and their effect on expression levels

TFBS turnover seems to be a common feature in orthologous enhancers of *Ciona* and *Phallusia* yet these enhancers maintain the same qualitative activity when studied by cross-species transgenesis. What remains unknown, however, is the extent to which these enhancers contribute to the different expression levels seen between these two species. Are expression level differences only due to unequal maternal factors? Or does the efficiency of the enhancers play a role too?

Further identification of the functional TFBSs by mutating them one by one would be necessary to confirm their role in the activation of the enhancer. Next, an analysis of their number and their affinity, compared between *Ciona* and *Phallusia*, could elucidate if there is a difference in efficiency between the homologous enhancers.

V. Shadow enhancers

Enhancer structure, such as TFBSs and spacer sequences, is not the only *cis*-regulatory feature controlling expression levels. It has also been shown that shadow enhancers controlling the expression of a gene within the same tissues or cells can also help control temporal or spatial expression robustness to genetic or environmental variability. However, there seem to be many different interpretations of what a shadow enhancer is. We first need to define a shadow enhancer.

1. Defining shadow enhancers

Eileen Furlong's lab defines them as "two enhancers that drive similar patterns of expression and in which deletion of one did not cause any obvious aberrant phenotypes" (Cannavò et al., 2016). This definition imposes several conditions. For the expression patterns to be just overlapping rather than completely identical, the pair of shadow enhancers would not necessarily share the same TFBSs. This principle was not the main focus when identifying the ascidian shadow enhancer for Brachyury that was identified based on the conservation of TFBSs between the pair of enhancers. Furthermore, in this paper, the timing of enhancer activity was ignored as a feature of shadow enhancers as they were both tested at mid tailbud stage even though initial Brachyury expression is at 64-cell stage, 7 hours earlier (Farley et al., 2016). As Brachyury is expressed for a long period of time throughout early embryogenesis in the notochord, it is not impossible to consider that different enhancers could take over the gene regulation at a different time point.

This raises the question of timing: do enhancers need to be active at the same time to be considered shadow enhancers? Or would it suffice that the enhancer is just has the potential to be active at the same time? Take for example enhancer competition driving *snail* in Drosophila that helps finely tune expression patterns and levels (Dunipace et al., 2011). To fulfil this function, neither shadow enhancer is redundant and therefore has an evolutionary purpose. Perhaps we should distinguish these different types of shadow enhancers and refer to a pair of enhancers with overlapping activity as distributed enhancers instead, as proposed by Scott Barolo (Barolo, 2012).

2. Evolution of shadow enhancer sequences

There are two schools of thought regarding shadow enhancers and evolution; a first that they promote evolvability due to the redundant nature of their activity (Cande et al., 2009; Hadzhiev et al., 2007; Hong et al., 2008; Paixão and Azevedo, 2010). This redundancy could be a source of evolutionary innovation by producing novel expression patterns. In this frame of mind, shadow enhancers would be rapidly evolving. A second is that, on the contrary, their function is far from redundant and they have a selective advantage making them even more conserved than other enhancers (Cannavò et al., 2016).

The collection of enhancers from this project and previous works within *Phallusia* and *Ciona* could allow further studies into enhancer sequence divergence. I would have been curious to compare the rate of evolution of enhancers to shadow enhancers within a genus and then compare the rate of evolution of orthologous enhancers between *Ciona* and *Phallusia*. This would firstly see if very fast evolving ascidians have more conserved shadow enhancers and secondly if orthologous enhancers are under the same selective pressures.

3. Selective advantage of shadow enhancers

If shadow enhancers are better conserved than other enhancers, this means that they bear a selective advantage. To test the role of shadow enhancers, I would have liked to test the phenotypic effect of removing one of the pair of shadow enhancers. There are three aspects that could be tested:

- Spatial expression precision: Do shadow enhancers promote precision of the expression pattern? By genome editing to remove one shadow enhancer, we could test if gene expression is affected. If the shadow enhancers maintain expression precision, removing one could cause expression to be lost in some cells or even just a slight loss in precision, which could be seen for instance by unilateral expression in certain cells.
- Temporal expression precision: There are actually two questions that we can tackle regarding timing. Firstly, do shadow enhancers increase responsiveness to activate a gene rapidly? Secondly, does the earliest enhancer need to be neighbouring the promoter to open the promoter? The first question would have to be addressed using shadow enhancers that do not neighbour the promoter. For the second, there are two pairs of shadow enhancers found during this project, regulating Otx and Eph1 genes, whereby one of the shadow enhancers is next to the promoter. Removing each enhancer, one at a time, by genome editing would be interesting to see how each shadow enhancers and their positions around the promoter can affect expression.
- Robustness to different conditions: Finally, one exciting experiment would be to test the importance of shadow enhancers in conserving development in response to variable environmental conditions. The shadow enhancers driving endodermal Eph1 expression could both be involved in gastrulation. If removing one of these enhancers still gave a viable embryo, it would be interesting to follow the effects of different temperatures on the gastrulation event.

Bibliography

Barolo, Scott. 2012. "Shadow Enhancers: Frequently Asked Questions about Distributed Cis-Regulatory Information and Enhancer Redundancy." *BioEssays: News and Reviews in Molecular, Cellular and Developmental Biology* 34 (2): 135–41. doi:10.1002/bies.201100121.

Cande, Jessica, Yury Goltsev, and Michael S. Levine. 2009. "Conservation of Enhancer Location in Divergent Insects." *Proceedings of the National Academy of Sciences of the United States of America* 106 (34): 14414–19. doi:10.1073/pnas.0905754106.

Cannavò, Enrico, Pierre Khoueiry, David A. Garfield, Paul Geeleher, Thomas Zichner, E. Hilary Gustafson, Lucia Ciglar, Jan O. Korbel, and Eileen E. M. Furlong. 2016. "Shadow Enhancers Are Pervasive Features of Developmental Regulatory Networks." *Current Biology: CB* 26 (1): 38–51. doi:10.1016/j.cub.2015.11.034.

Dunipace, Leslie, Anil Ozdemir, and Angelike Stathopoulos. 2011. "Complex Interactions between Cis-Regulatory Modules in Native Conformation Are Critical for Drosophila Snail Expression." *Development (Cambridge, England)* 138 (18): 4075–84. doi:10.1242/dev.069146.

Farley, Emma K., Katrina M. Olson, Wei Zhang, Daniel S. Rokhsar, and Michael S. Levine. 2016. "Syntax Compensates for Poor Binding Sites to Encode Tissue Specificity of Developmental Enhancers." *Proceedings of the National Academy of Sciences of the United*

States of America 113 (23): 6508–13. doi:10.1073/pnas.1605085113.
Femino, A. M., F. S. Fay, K. Fogarty, and R. H. Singer. 1998. "Visualization of Single RNA Transcripts in Situ." *Science (New York, N.Y.)* 280 (5363): 585–90.
Gualdi, R., P. Bossard, M. Zheng, Y. Hamada, J. R. Coleman, and K. S. Zaret. 1996. "Hepatic Specification of the Gut Endoderm in Vitro: Cell Signaling and Transcriptional Control." *Genes & Development* 10 (13): 1670–82.
Hadzhiev, Yavor, Michael Lang, Raymond Ertzer, Axel Meyer, Uwe Strähle, and Ferenc Müller. 2007. "Functional Diversification of Sonic Hedgehog Paralog Enhancers Identified by Phylogenomic Reconstruction." *Genome Biology* 8 (6): R106. doi:10.1186/gb-2007-8-6-r106.
Hong, Joung-Woo, David A. Hendrix, and Michael S. Levine. 2008. "Shadow Enhancers as a Source of Evolutionary Novelty." *Science (New York, N.Y.)* 321 (5894): 1314. doi:10.1126/science.1160631.
Iwafuchi-Doi, Makiko, Greg Donahue, Akshay Kakumanu, Jason A. Watts, Shaun Mahony, B. Franklin Pugh, Dolim Lee, Klaus H. Kaestner, and Kenneth S. Zaret. 2016. "The Pioneer Transcription Factor FoxA Maintains an Accessible Nucleosome Configuration at Enhancers for Tissue-Specific Gene Activation." *Molecular Cell* 62 (1): 79–91. doi:10.1016/j.molcel.2016.03.001.
Ledford, Heidi. 2017. "CRISPR Studies Muddy Results of Older Gene Research." *Nature News*. Accessed September 21. doi:10.1038/nature.2017.21763.
Paixão, Tiago, and Ricardo B. R. Azevedo. 2010. "Redundancy and the Evolution of Cis-Regulatory Element Multiplicity." *PLoS Computational Biology* 6 (7): e1000848. doi:10.1371/journal.pcbi.1000848.